The Dissertation Committee for Mohammad Hossein Rahmati
certifies that this is the approved version of the following dissertation:

THREE ESSAYS ON INDUSTRIAL ORGANIZATION AND INTERNATIONAL FINANCE

Committee:

David Sibley, Supervisor

Dean Corbae, Supervisor

Eugenio Miravete

Andrew Glover

Jacek Rothert

Matthias Kehrig

THREE ESSAYS ON INDUSTRIAL ORGANIZATION AND INTERNATIONAL FINANCE

by

Mohammad Hossein Rahmati, B.S.; M.S.Econ.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2012

To my beloved family; my supportive wife, my adorable daughter,

my devoted mother and my benevolent father

Acknowledgments

I am very grateful to my advisor David Sibley for guidance and encouragement. I am also thankful to Dean Corbae, Eugenio Miravete, Jacek Rothert, Seyed Reza Yousefi, Andrew Glover, Matthias Kehrig for their comments and suggestions.

THREE ESSAYS ON INDUSTRIAL ORGANIZATION AND INTERNATIONAL FINANCE

Mohammad Hossein Rahmati, Ph.D.
The University of Texas at Austin, 2012

Supervisors: David Sibley
Dean Corbae

This dissertation contains three chapters on industrial organization and international finance.

What motivates mergers in banking? The data show that merger activity is concentrated among very large banks. A large literature on the banking structure has studied this question by estimating cost functions and has provided mixed evidence. A crucial assumption is the exogeneity of input prices. If this assumption fails, result may be biased. This paper adopts the production function method proposed by [23] to separate the impact of productivity from scale economies in banking. To avoid this bias, I use recovery rates of non-performing loans, charge off rates, and cash holdings as proxies for productivity. The proxy method illustrates that the industry operates with significant diseconomies of scale, while the OLS method generates opposite results. Therefore, this finding supports the view that improvements in productivity cause mergers, which is also consistent with data. Finally, I introduce the Quantile Proxy Method to capture the impacts of both input endogeneity and

size heterogeneity. This method reveals that medium size banks have largest diseconomies of scale, while top 5% experience somewhat extensive economies of scale. This result sheds light on the fact why many mergers occur among large banks: large parties involved in a consolidation benefit from both productivity improvements and scale economies.

In the second chapter, we extend the work of [7] to a Small Open Economy. We introduce two additional wedges - a trend shock wedge and a debt price (country risk) wedge. We then evaluate the contribution of these wedges to the fluctuations in Mexico during the Tequila Crisis. Our results suggest that trend shock wedge is crucial to account for the behavior of the net exports and the current account during that time. Output movements are driven primarily by the traditional, stationary efficiency wedge. The role of debt price wedge appears to be minimal. On the theoretical side, we show that allocations in the prototype economy with a trend shock wedge are, up to a first-order approximation, identical to the allocations in a detailed economy with exogenous terms of trade shocks, when the latter are random walks. We provide evidence that in a few emerging economies terms of trade are well approximated by a random walk process. We also show that allocations in an economy with shocks to the world interest rate and to country spreads, coupled with a working capital friction, are identical to allocations in a prototype economy with debt price and labor wedge.

The main contribution of the third chapter is a simple theoretical framework and empirical estimations explaining the behavior of the manufacturers. The paper focuses on the frequently-used methods of demand estimation for discrete choice models to analyze the Iranian automobile market. It shows how both major companies in Iran choose to produce lower quality products and why they still charge high markups. Empirical estimations are based on [4] to predict marginal costs and markups. Estimation results also support the hypothesis that manufacturers are charging high markups. In addition, the counterfactual analysis carried out supports the view that both duopolist firms prefer to operate at lower quality rather than at higher quality production levels. We show that a decline in tariff rates may induce firms to produce high quality cars. Furthermore, analyses are performed using the Multinomial Logit methodology to better understand the Iranian automobile market. Tastes of people with different genders and ages for some specific cars are explained, and the effects of population changes on auto demand are predicted.

Table of Contents

Acknowledgments **v**

Abstract **vi**

List of Tables **x**

List of Figures **xi**

Chapter 1. Productivity, Scale Economies, Mergers: Estimating a Quantile Banking Production Function **1**

1.1 Background 6
1.2 Model 12
1.3 Estimation Result 18
1.4 Productivity 21
1.5 Merger 29
1.6 Quantile Model 33
1.7 Quantile Estimation Results 38
1.8 Conclusion 43

Chapter 2. Business Cycle Accounting in a Small Open Economy **45**

2.1 Introduction 45
2.2 Prototype Economy with Wedges 48
2.2.1 What is a wedge? 48
2.2.2 The prototype economy 49
2.2.3 Stationary prototype economy 53
2.2.3.1 Characterization 54
2.2.4 Wedges 55
2.3 The Accounting Procedure 56

2.4 Estimation and Accounting Results 60
2.4.1 Process for Wedges 61
2.4.2 Accounting for the Tequila Crisis 64
2.4.3 Contribution of wedges to fluctuations during the whole sample 67
2.5 Equivalence Results 68
2.5.1 A Detailed Economy with Terms of Trade Shocks 70
2.5.1.1 Are Terms of Trade Random Walks? 75
2.5.2 An Economy with Working Capital Constraint 76
2.6 Conclusions 78

Chapter 3. Demand Estimation for the Iranian Automobile Industry 79

3.1 Theoretical Model 81
3.2 Data 91
3.3 Empirical Models 94
3.3.1 Multinomial Logit 97
3.3.2 BLP(Berry, Levinsohn and Pakes) 98
3.4 Empirical Results 101
3.4.1 Multinomial Logit 101
3.4.2 The BLP Method 104
3.4.3 Counterfactual Experiment 105
3.5 Conclusion 110

Appendices 113

.1 Dynamic Theory Model 114
.2 Counterfactuals With Other Market Structure 115
.3 Multinomial Logit Estimation 117

Bibliography 119

List of Tables

1.1 Balance Sheets of Banks 7
1.2 Production Function for Financial Institutions. 20
1.3 Endogeneity of Inputs 22
1.4 Productivity Determinants 27
1.5 Distribution of Consolidation 29
1.6 Stylized Facts on Productivity Dynamics 32
1.7 Productivity Distribution among Mergers 33
1.8 Banks Input Portfolio 35
1.9 Quantile Banking Production Function. 39
1.10 Banking Production Function (Proxy Method) 40

2.1 Estimates of the matrix P 62
2.2 Estimates of the matrix Q 62
2.3 Correlation of simulated series with the data 70
2.4 Terms of Trade Random Walk Test 77

3.1 The Quality Game 83
3.2 Summary Statistics. 92
3.3 The Production and Market Share 93
3.4 Marginal effects for the Multinomial Logit Model 103
3.5 The BLP Estimation 105
3.6 Marginal costs and Markups for Automobiles 106
3.7 Profit in Quality Game 108
3.8 Profit for Counterfactual Quality Game 110
9 Profit in Quality Game (Other Market Structure) 116
10 Multinomial Logit Estimation 118

List of Figures

1.1 National HHI Index 9
1.2 Metropolitans HHI Index 10
1.3 Productivity Histogram 23
1.4 Productivity Dynamics 24
1.5 Productivity Dynamics During Merging 31
1.6 Banking Scale Economies 42

2.1 Estimated series for $\log \gamma_t$ and $\log q_t$. 63
2.2 Data and simulated series with all wedges 65
2.3 Mexican GDP during the Tequila Crisis. 67
2.4 Mexican net exports during the Tequila Crisis. 68
2.5 Mexican current account during the Tequila Crisis. 69
2.6 Terms of Trade in Emerging Economies 76

Chapter 1

Productivity, Scale Economies, Mergers: Estimating a Quantile Banking Production Function

Introduction

This paper focuses on three fundamental questions regarding the banking industry: First, what is the degree of scale economies in the banking industry? Second, has productivity growth increased after the deregulation of the mid 1990's? Third, how does consolidation between two banks affect productivity? The answers to these questions can shed light on what motivates banks to merge with each other. Banks often claim that these mergers help them to operate more effectively and to gain from scale economies.

It is important to know the answer to these questions to improve banking regulation. Policymakers often measure a bank's productivity by output per labor input. This way of calculating productivity often shows that productivity growth in banking lags behind productivity growth elsewhere in the economy. This measure is an incomplete gauge of efficiency, because banks can boost output per labor by investing more and by equipping workers with better tools in order to assess the best candidates for loans. That is, it is not clear that calculating productivity in this way can be compared in a meaningful way to other industries.

The problem of gauging productivity dates back to the seminal contribution, by [38] suggested a production function framework and defined technological change as the portion of output that is not explained by the amounts of inputs used in production. Moreover, he contended that the implicit load of assumptions is quite heavy in the conventional definition of productivity as output per labor, and his definition is considerably more general. I employ his approach by estimating a production function for banking. To the best of my knowledge, the "residual" approach to productivity has not been studied in the banking literature.

In applying this method in the context of banking it is necessary to properly classify inputs and outputs of the financial institutions consistently with the theory of the firm. According to [36], production in banking means a process that transforms borrowed funds (inputs) into loans and securities (outputs). They define the dollar volume of the various types of earning assets as the measure of the bank's output, which is analogous to a physical output of a manufacturer. The volume of earning assets is generally viewed as a stock variable from a portfolio standpoint; however, the continuing existence of balance sheet components requires continuing flows of services on the part of the banker to its consumers. In other words, as [30] states, earning assets and deposits are "not comparable to a stock of Rembrandt paintings but rather to a river, constantly renewed in the mountains and constantly disappearing down the valley, with the banker controlling the sluice." This notion of input and output is widely accepted in this literature.

Therefore, I define the total dollar value of loans and securities in a bank's balance sheet as its output. Similarly, labor, capital, and deposits are considered to be inputs into the financial firm's production process. Here, capital is defined by the value of equity capital in the bank's balance sheet and labor is the total compensation of employees. As inputs, I distinguish between checkable deposits and non-transactional deposits because in contrast to the latter, parts of the checkable deposits are subject to a reserve requirement.[1] Moreover, the ratio of these two deposits varies considerably with size, and this ratio can potentially play a crucial role in estimation.

Defining productivity as the portion of output that is not explained by the amounts of inputs leaves lots of room for it to account for heterogeneity in economic performances among banks. It is beyond the scope of this paper to describe all of the determinants of productivity in banking, but I will sketch out some basic ideas. One factor contributing to the growth of productivity is likely the creation of new types of deposits, each with better management of withdrawal timing and lower risk. For example, the expansion of certificates of deposit (CDs) and its link to commodity prices helps banks deal more effectively with risk and reduces the possibility of a bank run. Notice that in

[1]Money market deposit accounts (MMDAs), that introduced with the Depository Institutions Act in 1982, are included in checkable deposits, but not in the M1 definition of money thus not subject to reserve requirement

this example, the innovation in deposit types allows banks to produce more output with a fixed amount of deposits. This example is consistent with the above definition of productivity.

Another aspect of technological progress that improves productivity is the recent development in banking operations, which enables banks to fulfil the reserve requirement and create more loans even with no increase in capital. For instance, a financial institution can use new legal entities like Special Purpose Vehicles (SPEs) to isolate the entire company from a high risk project/asset. This transfer of loans from banks to SPEs is very common when banks issue mortgage-backed securities with payments coming from a pool of loans. Because capital requirements depend on the riskiness of bank assets, these entities tend to loosen reserve requirement constraints. Finally, the use of internet and computer, plus, improvements in the monitoring technology all should be counted as the rise in productivity.

Another focus of this research addresses scales economies in banking. In fact, a large literature in banking structure has studied this question and has obtained mixed results, although it generally contradicts the scale economy hypothesis except at a very low size level. Most of these studies use standard cost function estimation to examine scale economies, defined as the elasticity of cost with respect to output. These studies assumes that input prices are given, which rules out the possibility of the endogeneity of error terms. One noticeable example is [18], which provides solid evidence for the dependence of scale economies on the structure of banks capital and the intensity of diversification. Similarly, this paper attempts to capture the impacts of endogenous inputs that are plausibly related to productivity and estimates a production function, in contrast to a cost function method used by [18].

To illustrate how the endogeneity of inputs with respect to unobserved variables can bias the result and to put it more concretely, assume that the bank production presented by the Cobb-Douglas function, i.e. $Y_i = \Omega_i \prod_{j=1}^{J} X_{ji}^{\alpha_j} \zeta_i$. Here, the output (Y_i) of the bank i is produced by the multiple inputs (X_{ji}) plus the unobserved productivity (Ω_i). The economy of scale for this production function is simply defined as $\Sigma\alpha_j$, which is the variable of interest. Furthermore, the Duality Theorem illustrates that the cost function associated with this technology is:

$$log(cost_i) = \underbrace{\frac{1}{\Sigma\alpha_j}}_{\mu} log(Y_i) + \Sigma\frac{\alpha_j log(q_j)}{\Sigma\alpha_j} - \frac{1}{\Sigma\alpha_j} log(\Omega_i) - \frac{1}{\Sigma\alpha_j} log(\zeta_i)$$

Again, the inverse elasticity of cost with respect to output (μ) is called the degree of scale economies. As stressed by [18], the correlation of the unobserved variables (Ω_i) with the input prices (q_j) leads to biased estimation of scale economies. They put some structure on the problem by estimating a utility function for a manager who takes into account risks associated with production strategies as well as returns. In contrast, I will directly estimate the production function, and thus, I deal with the quantity data rather than the imputed price data, and I will use a proxy equation to capture the impact of unobserved variables

In Section 1.1, I describe some salient stylized facts about the banking industry and outline plausible theories for mergers. Section 1.2 reviews the proxy method introduced by [23] and discusses how it can be implemented to estimate a banking production function. Section 1.3 presents the results of the proxy method and shows the OLS estimates are biased. I will analyze the behaviour of productivity after the deregulation of mid-1990's in Section 1.4. Section 1.5 follows with a study of the productivity of the each party to the merger. Section 1.6 develops a method that is called the Quantile Proxy (QP) method to estimate a size-dependent production function. Section 1.7 presents the results of QP and Section 3.5 concludes.

1.1 Background

The banking industry is highly dynamic, with frequent mergers and acquisitions, and where fierce competition forces unproductive banks to exit the market. Table 1.1 shows that the number of active banks has shrunk rapidly, from around 11,000 banks in the mid 1990's to about 7,200 in 2007. Many factors contributed to this decrease. However, the data show that mergers are the dominant cause. According to FDIC statistics, there were over 6,340 mergers between 1993 and 2007 (an average of 422 per year). Only 111 of these mergers occurred when one party was at the brink of failure. In other words, more than 98% of mergers happened among two healthy banks with no assistance from the Federal Reserve. Therefore, a crucial question is what motivated those consolidations?

Table 1.1: Balance Sheets of Banks

Year	Bank	Consumer Loan	Business Loan	Real Estate Loan	Security	Capital	Labor	Checkable Deposit	Non-Transnational Deposit
1993	10,970	22.8	50.1	58.9	94.7	19.4	0.1	48.4	103.1
1994	10,475	26.2	54.2	63.6	104.6	20.6	0.1	50.8	104.5
1995	9,946	30.1	62.6	71.2	108.8	22.7	0.1	50.6	113.7
1996	9,539	32.5	70.6	76.2	112.2	24.7	0.1	49.4	126.9
1997	9,155	33.5	75.4	83.4	128.1	27.6	0.1	48.6	140.4
1998	8,803	34.0	87.4	91.7	143.2	31.0	0.1	48.7	157.1
1999	8,586	33.5	96.8	100.3	148.2	32.8	0.1	46.0	169.6
2000	8,317	35.6	103.6	114.1	156.3	35.2	0.1	44.3	184.7
2001	8,098	37.4	104.3	122.4	175.0	39.4	0.1	45.3	203.7
2002	7,890	41.3	100.4	136.2	192.3	43.8	0.1	45.9	224.7
2003	7,778	42.9	99.2	154.5	213.7	47.1	0.1	48.6	244.4
2004	7,637	46.9	103.1	171.3	230.1	53.8	0.1	49.2	263.0
2005	7,527	48.3	106.0	190.6	238.4	59.7	0.1	48.9	283.7
2006	7,407	47.2	114.3	207.8	253.8	64.5	0.1	45.7	301.8
2007	7,293	48.4	126.6	223.9	272.7	69.5	0.1	43.0	315.2

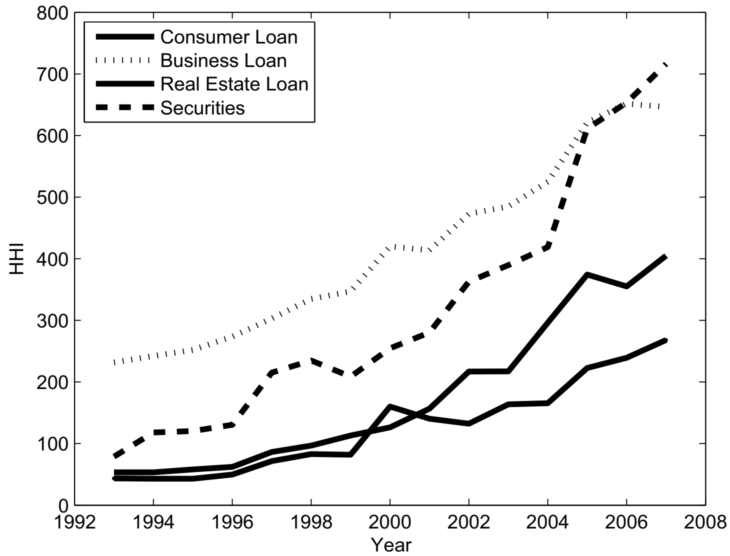
Numbers are unweighed average and in real term, 1983 equals 100

Table 1.1 demonstrates that the average bank size has steadily increased since 1993. For instance, in 2007 the average value of bank loans is three times the level it was in 1993. Similarly the average volume of deposits has increased about 2.4-fold during the same period. These facts may support the idea that market power encourages banks to consolidate. Indeed, looking at the market share of the very top banks indicates that a few giant banks have become dominant. For example in 2010, the top five banks hold more than 40% of total loans in the national market, while in 1993 only 17% of total loans originated with the top five banks. In the literature, one standard way economists use to examine concentration is to calculate the Herfindahl-Hirschman Index (HHI). Figure 1.1 plots the national HHI index based on the market shares of different assets: securities, real estate loans, business loans, and consumer loans. Interestingly, for all asset types the HHI indices have increased by at least 6 times; however, based on the Horizontal Merger Guidelines the banking industry is still categorized as an "unconcentrated market" with no sign of an excessive market power.[2]

Deriving the HHI from national data disregards the fact that retail banking is inherently a regional operation; that is, the commercial banks deliver services to local consumers. [15] document that depositors take into account distances to financial institutions in their utility, so that the cross-price elasticities between "close" banks are larger than those between "far"

[2]The Horizontal Merger Guidelines are published by U.S. Department of Justice and the Federal Trade Commission and is accessible through http://www.justice.gov/atr/public/guidelines/hmg-2010.html

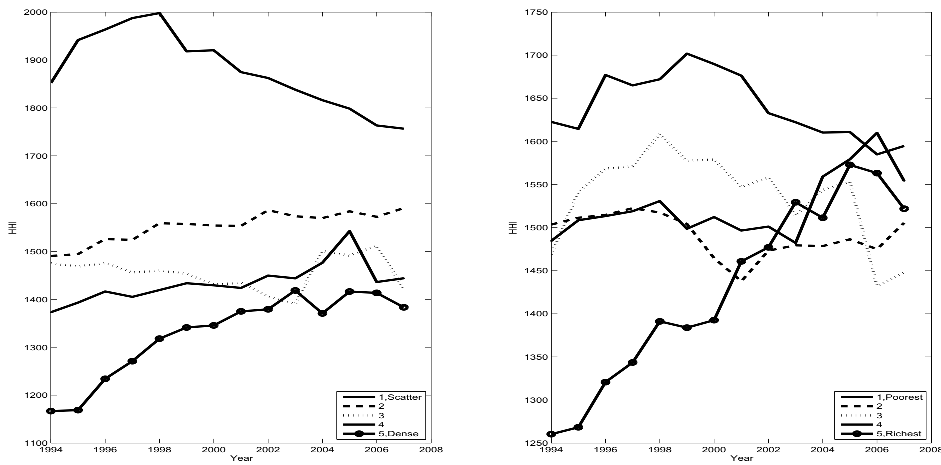
Figure 1.1: National HHI Index



banks. Therefore, I calculate the average HHI indices based on bank deposits for all metropolitan areas in the US. Figure 1.2 displays these HHI indices for different quantiles of income and population. Interestingly, these HHI indices are larger for metropolitan areas with scattered population and poorer neighbourhoods. This means that the larger the transportation cost or the smaller the demand, the higher is concentration. Given the rather moderate levels of concentration in banking, it seems unlikely that the main motivation for bank merger is the acquisition of market power. Therefore, I will seek elsewhere for a explanation.

A second candidate explanation is that merging banks seek economies of scale. As explained above, the economy of scale arises when costs of produc-

Figure 1.2: Metropolitans HHI Index



ing the same amounts of outputs decline after a consolidation. Many bankers report economies of scale as the main advantage of a merger. For instance, one year after the colossal consolidation of two giant banks: National Bank and Bank of America, the bank's annual report states "(because of) last year's historic merger $\cdots$ size and scale in key businesses allow us to offer customers and clients a wide choice of products and delivery methods at the lowest possible cost." [3] There is no way to evaluate these affirmations by just looking at data because revenues and costs are all endogenous variables, which are moved due to many factors unrelated to the merger. Therefore, answering to this question needs a banking framework to measure economies of scale, and this is what I intend to do in this paper.

[3]The full report can be downloaded from http://media.corporate-ir.net/mediafiles/irol/71/71595/reports/1998ar.pdf

A third explanation is that consolidation may lead to higher productivity; that is, banks operate more efficiently after the merger. As defined before, productivity is that portion of output that is not explained by the amounts of inputs used in production. In contrast, other papers, for example [25] focuses on the efficiency concept: defined as a gap between the actual and the optimal performance. These theories model the behavior of banks when they are maximizing profits or minimizing costs. My definition of productivity is free of any optimization hypothesis and depends on the assumptions on technology alone. Therefore,compared to other definitions, the productivity concept used in this paper may address more broadly the pros and cons of consolidation .

There is also some concern about how productivity is defined in this paper. For example, any amelioration in diversification or reduction in risk will be counted as an improvement in productivity. This change in productivity can be either a consequence of better risk management or just operating in a larger scale. As a result, I distinguish between "the economy of scale" and "scale effects". The former is a movement along the production frontier due to operating in larger sizes, while the latter is mainly a shift of the frontier because of larger scale. Some examples of these scale effects include: the consolidated portfolio of two bank with distinct specialities is less risky than each separate portfolio of them, or any interstate merger likely will lead to reduced risk as long as the income of the states are not perfectly correlated. In sum, I define "scale effects" as changes in "productivity" due to a change in size.

1.2 Model

First, I start with some assumptions about the functional form of production. Output, defined as the sum of loans and securities, is a function of inputs and productivity. As argued before, inputs are checkable deposits (c_t), non-transactional deposits (n_t), labor (l_t), and capital (k_t). Checkable deposits include bank accounts from which depositor can write a check to a third party. Non-interest bearing checking accounts, interest bearing NOW accounts, and money market deposit accounts are all examples. In contrast, non-transactional deposit accounts can not be used to write a check for others; and thus, they are offered with higher interest rates and with longer maturities. These accounts are divided into two types: saving accounts and time deposits (known as certificates of deposit). Table 1.1 shows that the share of non-transactional deposits have raised from 68% of total deposits in 1993 to 88% of total deposits in 2007. In other words, banks have shift their liabilities from short term to more manageable long term deposits.

Likewise, I define the labor input as total compensation of employees. The capital in bank's production function consists of two main components: physical capitals (like buildings and land) and equity capitals, which are reported in the bank's balance sheets.[4] Substituting these inputs into the Cobb-

[4]Physical capital is an tiny portion of total, so called, capital. For robustness check, when I include this component into non-transactional deposit, the final result doesn't change

Douglas technology and taking its logarithm lead to the following equation:

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_c c_{it} + \beta_k k_{it} + \beta_n n_{it} + \underbrace{\omega_{it} + \eta_{it}}_{\varepsilon_{it}} \tag{1.2.1}$$

This is the production function for the bank i, where all inputs are in the logarithmic form and t is a time index. Here, ε_{it} is the total error term composed of two terms: the bank's i unobserved productivity at the quarter t (ω_{it}) and an iid component denoted by η_{it}. The key difference of these two components is that ω_{it} is a state variable, and banks react to its variation. Therefore, estimating equation 1.2.1 using the OLS method may yield to a biased estimation of coefficients, if input variables are correlated with the unobserved productivity. As shown in Appendix II, the estimated coefficients of the OLS regression ($\hat{\beta}$) can be represented as a function of the true values of the coefficients (β) and cross correlations of right hand side variables (σ):

$$\hat{\beta}_l = \beta_l + \alpha_{ll}(\hat{\sigma}_{n,n} + \hat{\sigma}_{c,c} + \hat{\sigma}_{k,k})\sigma_{l,\omega} - \alpha_{lk}\hat{\sigma}_{k,l}\sigma_{k,\omega} - \alpha_{cl}\hat{\sigma}_{c,l}\sigma_{c,\omega} - \alpha_{ln}\hat{\sigma}_{n,l}\sigma_{n,\omega} \tag{1.2.2}$$

Where α_js are strictly positive parameters, and $\hat{\sigma}_{a,b}$ are cross correlations between the input a and the variable b. Although in a multivariate model, it is usually impossible to sign, per se, the bises of the OLS coefficients, the following analysis may highlight how the biases arise due to the endogeneity of inputs. If inputs are possitively correlated with each other (i.e. $\hat{\sigma}_{a,b} > 0$), which is the case in this industry and only labor responds to a productivity shock (i.e. $\sigma_{l,\omega} > 0$ and $\sigma_{k,\omega} = \sigma_{c,\omega} = \sigma_{d,\omega} = 0$), so based on equation 1.2.2 its OLS estimate is biased upward. On the other hand, because of the positive response of labor to productivity, the OLS underestimates the effects of capital and other inputs. The following equations underline this argument:

$$\overbrace{\hat{\beta}_l}^{\uparrow} = \beta_l + \alpha_{ll}(\hat{\sigma}_{n,n} + \hat{\sigma}_{c,c} + \hat{\sigma}_{k,k}) \overbrace{\sigma_{l,\omega}}^{+} - \alpha_{lk}\hat{\sigma}_{k,l} \overbrace{\sigma_{k,\omega}}^{0} - \alpha_{lc}\hat{\sigma}_{c,l} \overbrace{\sigma_{c,\omega}}^{0} - \alpha_{ln}\hat{\sigma}_{n,l} \overbrace{\sigma_{n,\omega}}^{0}$$

$$\underbrace{\hat{\beta}_c}_{\downarrow} = \beta_c + \alpha_{cc}(\hat{\sigma}_{n,n} + \hat{\sigma}_{l,l} + \hat{\sigma}_{k,k}) \underbrace{\sigma_{c,\omega}}_{0} - \alpha_{ck}\hat{\sigma}_{k,c} \underbrace{\sigma_{k,\omega}}_{0} - \alpha_{cl}\hat{\sigma}_{l,c} \underbrace{\sigma_{l,\omega}}_{+} - \alpha_{cn}\hat{\sigma}_{n,c} \underbrace{\sigma_{n,\omega}}_{0}$$

In sum, the simultaneity problem in the standard OLS method leads to a biased estimation of the coefficients. It is worth mentioning that a large body of banking literature have performed the OLS method to estimate cost functions. As a result their findings may be affected by this simultaneity problem.

To solve the simultaneity problem, an alternative approach is proposed by [29] and [23]. In general, they introduce a structural equation along with equation 1.2.1, which contains a proxy variable as well as productivity. For example, Olley-Pakes use an equation that write investment as a function of the current level of productivity and the state of capital. Therefore, in this example the investment acts like a proxy for productivity because given the state of the capital the higher the productivity, firms invest more. Similarly, Levinsohn-Petrin consider the demand for a freely variable input, for instance electricity usage. Next, in the same spirit, they introduce the intermediate input's demand function as an structural equation, which again is a monotonic function of productivity and also depends on other fixed state variables. In sum, the Levinsohn-Petrin method uses the demand of a intermediate input as a proxy for the firm's productivity.

In the same way, I apply a proper modified version of their proxy approach for financial institutions. In banking, neither the investment decision nor the intermediate input demands is an applicable candidate for a proxy in banking, mainly because of lack of data for physical capital investment and demand of an intermediate input like electricity uses. Alternatively, I argue that either recovery rates or cash holdings can be employed as a proxy for productivity. To illustrate why banks with higher productivity required to hold less amounts of cash in their balance sheet, consider banks with better timing management of deposits. Empirically, these institutions hold high quality deposits with settled timings of withdrawal. Thus, they only need to hold a minimum level of cash, with no risk of runs. Compared to saving accounts that can be withdrawn at any time, CDs which generally have a fixed maturity length can be rated as a high quality deposit, albeit one that costs more. Therefore, the productivity that arises from X-efficiency in management allows banks to have less cash in their balance sheets.

To set up a structural model for cash holdings, some institutional knowledge is necessary. First, based on the above discussion the cash holding equation is a decreasing function of productivity. Banks decide on their required cash reserves based on their past amounts of checkable and non-transactional deposits. Ideally, banks take into account maturities of these deposits; in particular, past deposits that mature in this period. Once again, because of limitations in data I use deposits lagged one quarter. Finally, because the structure of the bank is also a determinant factor in the cash holding, I employ lagged amounts of cash as another explanatory variable. Therefore, I set up the structural cash holding equation as following [5]:

$$s_{it} = S(\omega_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1}) \tag{1.2.3}$$

Here, s_{it} is the value of cash in the bank's i balance sheet at quarter t. The next step is to obtain productivity from equation 1.2.3. The assumption that S is a monotonic function of its first argument allows me to invert the function S and to derive the productivity as a function of other variables: $\omega_{it} = \Omega(s_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1})$.[6] Replacing this equation into equation 1.2.1 results in:

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_c c_{it} + \beta_k k_{it} + \beta_n n_{it} + \Omega(s_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1}) + \eta_{it} \tag{1.2.4}$$

Defining Φ as $\Phi(s_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1}) = \Omega(s_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1}) + \beta_n n_{it}$ simplifies equation 1.2.4 and makes it ready for estimation:

$$y_{it} = \beta_0 + \beta_l l_{it} + \beta_c c_{it} + \beta_k k_{it} + \Phi(s_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1}) + \eta_{it} \tag{1.2.5}$$

[5]In the Appendix III, I introduce two other proxies for productivity in the banking industry: "charge-off rate" and "recovery rates of non-performing loans". These proxies are a tad different from the standard proxy, because the standard proxy represent an structural decision equation, while both "charge-off rate" and "recovery rate" are like exogenous functions of productivity. Their structural equation are :

$$\begin{aligned} Charge-off_{it} &= C(\omega_{it}, n_{it}, Charge-off_{i,t-1}) \\ Recovery_{it} &= R(\omega_{it}, n_{it}, Recovery_{i,t-1}) \end{aligned}$$

[6]In my empirical estimation for this equation, productivity is monotonic with respect to cash holding in the region $\big(\bar{S}_t - 4std(S_t), \bar{S}_t + 4std(S_t)\big)$, when other variables in the equation 1.2.3 are at their average levels. Here, $\bar{S}_t$ and $std(S_t)$ are respectively the average and the standard error of cash holding. Interestingly, the relationship is also negative for this example.

Importantly, the assumption that η_{it} are iid random variables ensures that all regressors in equation 1.2.5 are orthogonal to error terms. Therefore, in the first stage an estimator that is linear in l, c, k and non-parametric in Φ arguments can be used to obtain consistent estimates of $\beta_l, \beta_c, \beta_k$. Following Levinsohn-Petrin, I use the OLS to estimate equation 1.2.5 with a third-order polynomial in s_{it}, n_{it}, $n_{i,t-1}$, $c_{i,t-1}$, and $s_{i,t-1}$ to approximate $\Phi(.)$. Notice that it is crucial to get a consistent estimates of β_l, β_k, and β_c, because they will be used in the second stage. Moreover, given the result of this estimation I can calculate Φ_{it} for all banks.

In the second stage, because $\Omega(.)$ has a linear term of n_{it}, this term appears twice in the function Φ, and another assumption is required to estimate β_n. Olley-Pakes assume that productivity follows a first-order Markov process, hence its expectation can be calculated by a linear operator $E[\omega_{it}|\omega_{i,t-1}]$. Under this assumption, I can set up the second stage of estimation as:

$$\begin{aligned} \min_{\beta_n} \quad \sum_t \sum_i & \quad (\zeta_{it})^2 \\ \omega_{it} &= \hat{\Phi}_{it} - \beta_n n \\ \zeta_{it} &= y_{it} - \hat{\beta}_l l_{it} - \hat{\beta}_c c_{it} - \hat{\beta}_k k_{it} - \beta_n n_{it} - \hat{E}[\omega_{it}|\omega_{i,t-1}] \end{aligned} \tag{1.2.6}$$

Where $\hat{\beta}$ comes from the first stage, and $\hat{E}[\omega_{it}|\omega_{i,t-1}]$ is an expectation operator given the Markov assumption.[7] This estimator produces a consistent

[7]For example it can be represented by

$$\hat{E}[\omega_{it}|\omega_{i,t-1}] = \alpha_0 + \alpha_1\omega_{i,t-1} + \alpha_2\omega_{i,t-1}^2 + \alpha_3\omega_{i,t-1}^3$$

estimates of β_n because ζ_{it}, that is the residual of the productivity process, and η_{it} are orthogonal to n_{it}. Moreover, the calculated productivities (line 2 of the equation 1.2.6) are consistent and will be used in Section 1.4 and 1.5 to study their dynamics. Finally, to measure the precision of coefficients I employ the bootstrap method because calculating the standard error is so hard in this two step approach. Two resampling procedures are applied in this paper: resample banks and take all records belong to these banks, or resample all observations. Next, for each sample all parameters are estimated using the proxy method explained above. Given the distribution of bootstrap parameters, standard errors can be computed easily.

1.3 Estimation Result

' This section reports the coefficients of the standard method (OLS) and of the proxy approach to highlight the importance of dealing with endogeneity in freely variable inputs. Table 1.2 shows the results of estimating a production function for financial institutions using the standard method and the proxy approach with different proxies. The first column depicts the coefficients from the OLS regression. It shows moderate scale economies for banks and displays a large share of capital in production. In particular, this estimates imply that capital is solely responsible for about 45% of production, but Table 1.1 documents that capital accounts for an average of 10% of total bank liabilities.

Where α's are the result of OLS estimation. Notice that α's depends on the value of β_n and thus for each iteration must be recalculated.

This discrepancy can be attributed to the role of capital in banking. Capital serves more than just provides liabilities for loans. The more capital lowers the risk of banking, so it provides assurance to make more. Therefore, it is intuitive that the share of capital in production is higher than its portion of liabilities.

Moreover, a coefficient in the production function can be interpreted as the share of the corresponding input in total costs. Based on official reports the employee salaries and benefits on average account for about 30% of total interest and non-interest expenses.[8] Therefore, an estimate of $\beta_l = 0.15$ determines that the role of labor in production is less than its expense share. However, one might argue that accounting reports understate the share of capital when calculating interest expense. Nevertheless, comparing inputs' various coefficients can reveal their respective weights. For instance, in aforementioned official reports total interest deposit expenses account for about 33% of total expenses while my estimates shows a share of around 36%. In sum, the standard method highlights the weight of capital and plays down the role of labor in production.

Section 1.2 explored the possibility that the bias in the OLS estimate of coefficients is caused by the endogenous inputs choice of bankers in response to productivity shocks. One possible solution to this problem is the proxy method. Table 1.2 shows the results of the proxy method with different

[8] Table CB07 in http://www2.fdic.gov/hsob

Table 1.2: Production Function for Financial Institutions.

Variable	OLS	Proxy Method		
		Proxy: Cash	Charge-off Rate	Recovery Rate
labor	0.157***	0.090***	0.100***	0.085***
	(0.002)	(0.002)	(0.003)	(0.002)
checkable deposit	0.080***	0.090***	0.100***	0.109***
	(0.001)	(0.004)	(0.003)	(0.005)
non-transactional deposit	0.352***	0.358***	0.400***	0.4220***
	(0.003)	(0.012)	(0.020)	(0.018)
capital	0.457***	0.312***	0.348***	0.328***
	(0.003)	(0.003)	(0.005)	(0.004)
scale economies	1.046***	0.879***	0.950***	0.946***
	(0.002)	(0.012)	(0.015)	(0.010)

Standard errors in parentheses, *** p<0.01, ** p< 0.05, * p<0.1

Dependent variable is output that is sum of loans and securities.

proxies: cash holdings, charge-off rates, and recovery rates of non-performing loans. Importantly, the labor coefficient is smaller in all proxy estimations than the result of the standard OLS model. The decline in the labor effect is in line with the finding of Levinsohn-Petrin. Their main argument is that labor is more responsive to technology shocks and adjusts more freely compared to other inputs. As a result, in a method that ignores productivity in its estimation, the effects of productivity on output enter in the labor coefficient. Therefore, because of positive correlation of labor and productivity, the OLS overestimates the labor coefficient.

Significantly, the two approaches have dissimilar outcome for scale economies. The standard model displays moderate scale economies, where the proxy model shows small diseconomies of scale. I note that even with diseconomies of scale, merging parties may benefit from increased scale if that makes it easier to achieve efficiencies.

How does one test for the endogeneity of inputs and their dependence on productivity? For this purpose, table 1.3 documents the OLS regression of inputs one by one on productivity and other inputs. As is clear from this table, all inputs would significantly depend on the productivity, which confirms why OLS yields to a biased estimation. Furthermore, the larger the coefficient of the productivity, that the greater the biases of the OLS. This illustrates why in table 1.2 the highest difference between the OLS and the proxy method appears in the capital and the labor coefficients. In the next section I will investigate on other aspects of productivity behavior.

1.4 Productivity

This section studies and compares the productivity estimates derived by the above methods. It is worth mentioning that the productivity in the standard method is the same as the Solow residual in the real business cycle literature, while the productivity in the proxy method is calculated by inverting the proxy equation. Figure 1.3 depicts the histogram of productivity for years 1993 and 2007. The right digram graphs the result of the OLS method and the left one shows the result of the proxy method using the cash holding

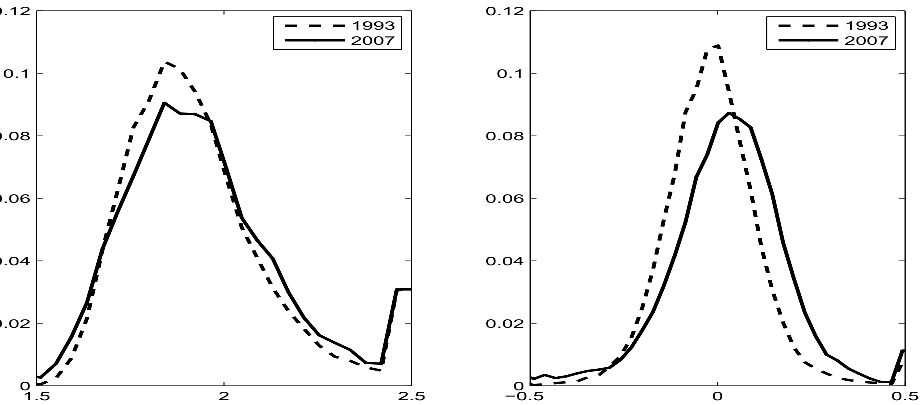
Table 1.3: Endogeneity of Inputs

	Labor	Capital	Checkable Deposit	Non-transactional Deposit
productivity	0.868***	1.653***	-0.366***	-0.571***
	(0.005)	(0.004)	(0.007)	(0.005)
capital	0.199***		0.084***	0.685***
	(0.001)		(0.002)	(0.001)
checkable	0.317***	0.032***		0.081***
	(0.001)	(0.001)		(0.001)
non-transactional	0.283***	0.479***	0.151***	
	(0.001)	(0.001)	(0.002)	
Labor		0.175***	0.737***	0.355***
		(0.001)	(0.002)	(0.001)
Constant	-8.165***	-3.396***	5.189***	4.905***
	(0.010)	(0.013)	(0.022)	(0.016)
Observations	533,329	533,329	533,329	533,329
R-squared	0.887	0.908	0.760	0.880

Standard errors in parentheses, *** p<0.01, ** p< 0.05, * p<0.1

Regression of inputs on the calculated productivity and other inputs

Figure 1.3: Productivity Histogram



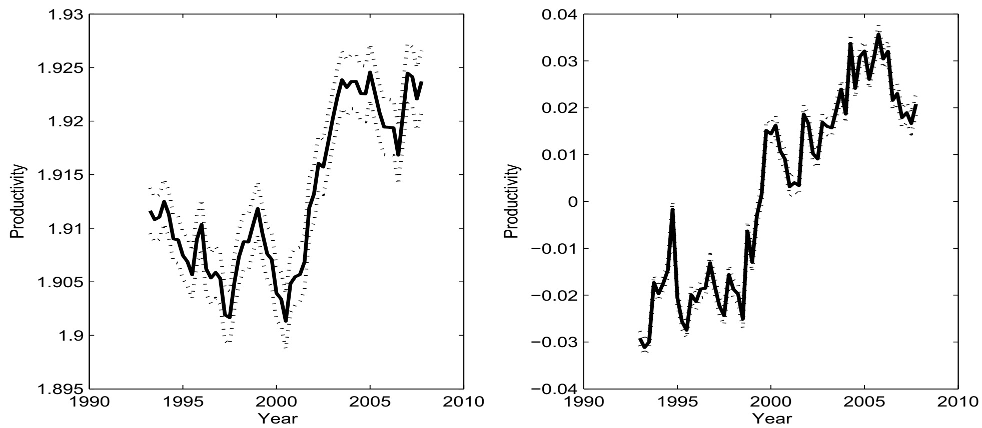
Histogram distribution of productivity in 1993, 2007, right graph is the result of the OLS estimation and left graph depicts the proxy method result. Vertical axis is the frequency.

as the proxy equation.[9]

Figure 1.3 implies that the kurtosis of the productivity distribution has decreased about 15% in recent years compared to 1993. In other words, the mass of firms in the middle of the distribution and the peakedness of productivity histogram have dropped. Interestingly, the histograms for 2007 and 1993 illustrate that the thickness of the right tail is unchanged. Therefore, a heavier tail hasn't caused the decline in the peakedness of the profile. Moreover, the average productivity remained unchanged between 1993 and 2007, but the standard error has slightly decreased. Finally, as is clear in the right panel of Figure 1.3 the productivity distribution shifted toward the right in 2007 compared to 1993, and a standard distribution test demonstrates that

[9]The diagrams for other proxies are similar to these results

Figure 1.4: Productivity Dynamics



Evolution of average financial institutions productivity between 1993-2007, right panel is OLS result and left panel draws the proxy method result. Vertical axis is log productivity.

log-normal distributions can best describe both graphs.

In banking, productivity growth has long been the focus of government reports on the industry. Typical findings support the view that the productivity grows at around 1.5% percent annually, compared to 4% annual growth for manufacturing industries. In contrast, the models used in this paper predict less than one percent annual growth of the productivity.

As is depicted in Figure 1.4, the OLS result shows steady growth for the banks productivity between 1993-94, then a sharp decline on the first quarter of 1995 continued with "wobbling" growth until 2006, with subsequent deterioration. In contrast, for the mid-nineties the proxy estimates display a deterioration in banking productivity during 1993-1997, which is in line with findings reported by [3]. Strikingly, the proxy model finds rapid productivity growth in a period shortly before the end of the 2001 crisis until 2003Q3, with an average annual growth rate of 12%. For the rest of the period, the proxy results indicate very low productivity growth in banking.

The next step is to provide some insights into the structural determinants of the variation in productivity among banks. Previous research has suggested that optimal scale, scope, and diversification can improve efficiency in banking ([25]). To better understand how size influences productivity, consider the technological adoption of two very different banks: one with over 1000 branches all across the country, and one small financial institution with a single branch.[10] In this environment a new innovation such as a new internet tool, introduced in the market. The large bank will employ this technology extensively in its operation to facilitate the connection between branches and provide better financial services for consumers, and to form inter-bank relationships. In contrast, the small bank never needs to use this product and never gains from this innovation. Although, not the same as the normal definition of economies of scale, this efficiency is definitely related to scale.

With respect to economies of scope, if there are productive complementarities between outputs, then producing these products is efficient. For example, monitoring real estate loans and consumer loans requires an extensive collection and analysis of data. Obviously, detailed knowledge of consumers'

[10]Notice that a bank with just one branch is not a hypothetical assumption. In 1993 about 40% of commercial banks operate with a unit institution. This share is still high in 2007 and around 27% of all banks

financial histories helps banks to improve their assessments of the credit worthiness of mortgage loans. Therefore, having both types in the asset portfolio may provide an information processing advantage. Similarly, a diverse pool of complement securities allows banks to reduce risks and transaction costs. In addition to scale and scope, it is obvious that a risk reduction caused by a diversification achieves to higher productivity.

To evaluate the impact of these factors on productivity, I regress the productivity estimates from the OLS and the proxy method son a set of explanatory variables. There are: size, index of diversification, share of loans in total assets, and other variables that are shaping the structure of a bank. To construct the diversification index, I update the index proposed by [17], which is gauged by a bank's exposure to macroeconomic risks. It is based on the idea that the more geographically diverse the branches, the bank is less vulnerable to macroeconomic risks.

To construct the diversification index, I first compute a variance-covariance matrix, $\mathbf{V}$, of unemployment rates in all states over the period 1993-2007. Next, I calculate the portion of state j deposit in bank i, at quarter t as s_{ijt} with vector $S_{it} = [s_{1jt}, \cdots, s_{50jt}]$. These deposit shares serve as weights for the diversification indices defined by $1/(S'_{it}VS_{it})^{1/2}$. This index states that If a unit-institution bank operates in a state with a low variance, then it has a higher diversification index compared with a bank in a state with high variance. Similarly, entering a new state increases the bank's diversification index because the economic performance of the new state likely is not perfectly cor-

Table 1.4: Productivity Determinants

	Proxy			OLS		
size	0.161***	0.160***	0.168***	0.020***	0.019***	0.067***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
diversification	0.010***	0.016***	0.014***	0.041***	0.047***	0.042***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
portfolio concentration		0.201***	0.184***		0.217***	0.147***
		(0.002)	(0.002)		(0.002)	(0.002)
branch number			-0.008***			-0.041***
			(0.000)			(0.001)
branch spread			0.006***			0.007***
			(0.000)			(0.000)
house			0.012***			0.017***
			(0.001)			(0.002)
new house			0.000			0.001***
			(0.000)			(0.000)
population			-0.015***			-0.031***
			(0.001)			(0.002)
establishment			-0.007***			-0.032***
			(0.001)			(0.001)
employment			0.005***			0.018***
			(0.000)			(0.001)
age			-0.050***			0.056***
			(0.002)			(0.003)
Constant	1.726***	1.650***	1.802***	-0.002***	-0.084***	-0.189***
	(0.000)	(0.001)	(0.008)	(0.001)	(0.001)	(0.013)
observations	495,911	495,911	486,645	497,246	497,246	487,916
R-squared	0.783	0.790	0.810	0.023	0.037	0.097

Standard errors in parentheses, *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Regression of calculated productivity on size, diversification, portfolio choices and other variables of bank's structure. Explanatory variable in the left (right) panel is calculated productivity from proxy method (ols).

rolated with the current bank's portfolio. In contrast to [17] that used this definition for one year, my indices vary over time because of many interstate consolidation after 1997.

I define the portfolio concentration as the sum of the square shares of each output in the portfolio. Thus, the portfolio concentration measure is defined as $\sum_{i=1}^{4} s_i^2$ for a bank with four types of output: consumer loans, business loans, read estate loans, and securities. In table 1.4 the branch number is counts of branches operating under a bank's title. The branch spread is an average distance (in miles) of the bank's branches from its headquarter. Other structural variables in Table 1.4 are calculated by using demographic data of branch locations. For instance, the house variable is an average number of houses in the zip codes where a bank's branches are located.

Among other things, the interesting result visible in Table 1.4, is the impact of size and diversification on productivity: controlling for every other variable the larger the size of a bank, the higher its productivity. In addition, the size coefficient is large in magnitude compared to other variables. This result suggests that banks gain from improving productivity because of operating in a larger size, although they may not benefit from economies of scale. This effect is called scale efficiency, compared to economies of scale, and it shift the production frontier in the same way as other efficiency does.

Another key result is that diversification improves productivity with a much smaller magnitude compared to size but still significant. For example, in the proxy regression a one-standard-deviation increase in the size raises

Table 1.5: Distribution of Consolidation

		acquired bank					
	size rank	1	2	3	4	5	sum
	1	0.025	0.002	0.000	0.002	0.007	0.036
acquirer	2	0.032	0.022	0.006	0.007	0.017	0.084
bank	3	0.044	0.027	0.016	0.011	0.030	0.128
	4	0.044	0.035	0.032	0.022	0.090	0.223
	5	0.026	0.042	0.074	0.092	0.291	0.525
	sum	0.171	0.128	0.128	0.134	0.435	

Share of each pair of size-rank in total consolidations between 2000-2007

the productivity by 0.9 standard deviations, and the same exercise for the diversification improves the productivity by 0.01 standard deviations. These positive impacts also hold for the OLS results, and again the size impact is more pronounced. Surprisingly, portfolio concentration promotes productivity. That is, this exercise provide no evidences for economies of scope.

1.5 Merger

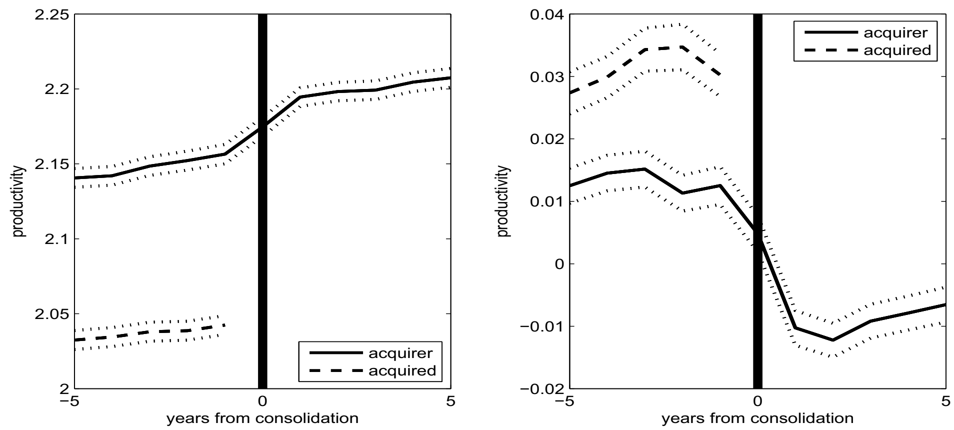
What motivates banks to consolidate? As demonstrated above, many studies argue that two factors cause consolidation in banking: improvements in efficiency and economies of scale. Providing evidence to assess these causes requires setting up a model to measure both efficiency and economies of scale. In other words, there is no direct evidence in the data that can support which one is the main cause of consolidation. However, some stylized facts may prove useful.

Table 1.5 reports the distribution of consolidation based on the sizes of parties. I rank financial institutions by their asset values in five equal quantiles: the smallest banks in group 1 and the top 20% banks in group 5 [11]. Then, I counted numbers of mergers in each pair-group to compute the share of each pair in the total consolidations. Importantly, Table 1.5 shows that about one third of the consolidations occur among top 20% of banks. Section 1.7 attempts to answer the question why large banks are more likely to merge? Significantly, Table 1.5 shows that acquirer banks are on average larger than acquired banks. Consequently, it may therefore imply that the acquirers are more productive than acquired banks because productivity increases by size, as documented before.

Figure 1.5 shows the average productivity of acquirer and acquired before and after a merger. Improvement in productivity can be proposed as a primary cause of consolidation if after the merger either the productivity growth rises or the less efficient bank performs as well as the more productive one. Noticeably, the proxy method predicts that productivity improves in the short run aftermath of a consolidation. In contrast, an opposite picture is depicted by the OLS result, where after a consolidation the productivity deteriorates both in the short run and in the long run. In addition, the Proxy model suggest that acquirers are on average more productive than acquired banks, while an opposite is reported by the OLS.

[11] Because numbers of financial institutions are decreasing during years, numbers of banks in each group is different in each year. That said, on average numbers of institutions in each group is about 2000 banks.

Figure 1.5: Productivity Dynamics During Merging



Productivity of banks before and after a consolidation, the right (left) panel is the OLS (proxy) result

The discussion of Table 1.5 points out that because productivity increases with size, we expect that banks emerged from consolidation become more productive. This view is consistent with the explanation provided by the proxy estimation. However, It is natural to ask whether it can be confirmed by other evidence directly from the data. Table 1.6 shows two other patterns in banking. I focus on two variables in the data that can serve as a basis for productivity: income per asset and asset growth. For instance, the higher the productivity, the larger both the income per asset and the asset growth of the bank. Interestingly, both indices show an increasing trend after consolidation, which is consistent with the proxy results. Moreover, these evidences are again in accord with the proxy results in which acquirers are more productive than acquired banks.

Table 1.6: Stylized Facts on Productivity Dynamics

pre-merger			post merger
	acquired	acquirer	new bank
productivity proxy	2.03	2.14	2.21
std	0.01	0.01	0.01
productivity OLS	0.03	0.01	-0.01
std	0.00	0.00	0.00
income per asset	200.42	240.35	257.76
std	4.90	13.12	4.58
asset growth	2.54%	4.53 %	5.09 %
std	1.02%	0.79%	0.86%

Average of variables two years before and after acquisition.

Another interesting comparison is between the distribution of consolidation produced by each method. Table 1.7 shows the distribution of merger conditional on size. The left panel (proxy outcome) illustrates that 25% of mergers occur within the top-quantile banks, but this probability is quite small for the non-large banks. In contrast, the OLS result (right panel) depicts close to a uniform distribution with respect to size; that is, scale contributes equally for all size banks.[12] As before, the proxy result is in line with the fact that the willingness to merger is more pronounced among larger banks. To see this, remember that productivity is positively correlated with size, and if large banks consolidate more often, the distribution of mergers must be more skewed toward high productivity levels.

In sum, the proxy model rules out scales economies as a key explana-

[12]Likewise, the quantile estimation in the next section support this conclusion

Table 1.7: Productivity Distribution among Mergers

		acquired bank									
		proxy					OLS				
	quantile	1	2	3	4	5	1	2	3	4	5
	1	0.029	0.008	0.008	0.007	0.025	0.061	0.048	0.041	0.037	0.042
acuirer	2	0.020	0.011	0.011	0.012	0.044	0.055	0.041	0.035	0.028	0.039
bank	3	0.026	0.028	0.019	0.020	0.049	0.052	0.040	0.034	0.031	0.042
	4	0.030	0.033	0.036	0.030	0.073	0.036	0.038	0.036	0.034	0.038
	5	0.012	0.035	0.068	0.094	0.271	0.028	0.028	0.031	0.038	0.067

Share of each pair of productivity-rank in total consolidations between 2000-2007, Right (left) panel is for OLS (proxy) model

tion for bank mergers; however, it provides evidence that all parties benefit from the productivity growth and improvements in efficiency. This result is consistent with the fact that banks experience rapid asset growth and a higher income per asset after a merger. Moreover, the proxy model depicts a skewed distribution of mergers toward high productivity levels. In contrast, the OLS model highlights the role of scale economies and shows a deterioration in productivity after a merger. It displays a uniform distribution of merger across banks with different productivity levels.

1.6 Quantile Model

One concern that has been raised frequently both among empirical economists and within policy makers, is whether technology used by large banks differs from what employed by small banks. The difference in the technology can be attributed to a number of causes, including institutional factors and banking structures. For instance, in recent years it has been argued over and over that large banks have been strengthened by regulation because they are subject to too big to fail, and their connections and lobbies help them to secure their equities in the market and benefit from government programs (like Troubled Asset Relief Program (TARP))

Moreover, many papers contend that large banks operate with different organizational structures compared to small banks. For instance, [39] discusses that small banks widely use "soft" information that can not be credibly transmitted. Thus, a unit institution structure with single manager is more appropriate for very small banks. In contrast, large banks generate and use "hard" and verifiable information to pass along inside the banks, which performs better in a large hierarchical institution. [3] provides evidences for the different information technologies used by large and small banks. In particular, they show that bank-firm relationships tend to be stronger for small banks. In contrast, large banks frequently lend to larger firms with farther away distances from them.

As is shown in Table 1.8 the structure of the bank's balance sheet are apparently dissimilar for different sizes. Although the input portfolio is an endogenous decision, this table demonstrates that capital constitutes about 17% of the total input values for very large banks, around 10% for middle banks, and for very small banks more than 40%. This U-shaped profile of the capital share adjusts by more utilization of non-transactional deposit. In other words, as banks get larger, they can attract additional large denomination deposits

Table 1.8: Banks Input Portfolio

Input	low 1%	low 10%	low 50%	top 50%	top 10%	top 1%
Capital	44.0	18.7	13.3	11.1	13.7	17.3
Labor	0.2	0.1	0.1	0.1	0.1	0.1
Checkable Deposit	21.0	24.9	25.1	21.6	15.4	15.7
Non-transactional Deposit	34.5	56.1	61.4	66.6	70.9	66.6

The share of inputs in banks' balance sheets for different bank sizes.

with longer maturities. These structural shifts underline the importance of studying a size dependent technology for policy making, i.e. the interbank regulations.

Many empirical papers on banking structure also highlight the importance of size in scale economies. [16] stress that the scale economies can vary among banks with different size, and in contrast to earlier studies, they show that large banks benefit from considerable scale economies. Interestingly, they document that standard cost function estimation fails to show scale economies for large banks, while their method provides an opposite conclusion because it can control for trade-offs between risk and return associated with different investment strategies. That is, capturing the endogeneity of risks entails an estimation with considerable scale economies for very large banks. The aforementioned facts and these findings support the view that the proxy approach may lead to a wrong conclusion for very large banks and it deserves more scrutiny.

To address this issue in the context of the production function estimation, I extend the Levinsohn-Petrin method by introducing a two step quantile regression. It is useful to remember that in the standard proxy model, the intermediate input demand is assumed to be a monotonic function of productivity, as a result unobserved variables can be constructed by inverting this demand system (Equation 1.2.3). Despite its straightforward intuition, a crucial implicit restriction is imposed during the estimation by assuming that the input demand of all firms can be modeled by an identical equations. It can bias the results, in particular when firms behave differently as they become larger.

The homoheneity assumption is imposed when a linear OLS regression is used to estimate the production function. It can be relaxed to a weaker assumption by employing a quantile approach, where the monotonicity of the demand system with respect to productivity must hold as a core assumption. Indeed, the quantile approach is built based on the assumption that banks with the same size using a similar technology. Importantly, the results of the quantile method can be exploited to construct a test for the homogeneity assumption imposed by the OLS regression.

Similar to the proxy approach, I can rewrite equation 1.2.6 as following.

$$y_{it} = \beta_0^\tau + \beta_l^\tau l_{it} + \beta_c^\tau c_{it} + \beta_k^\tau k_{it} + \Phi^\tau(s_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1}) + \eta_{it}^\tau \quad (1.6.1)$$

In this equation, Φ^τ is a third-order polynomial of its arguments and $\beta_l^\tau, \beta_c^\tau, \beta_k^\tau$ are consistent estimates of the input coefficients for quantile τ. To apply this method [19] introduces the check function defined by $\rho_\tau(\eta) = \eta(\tau - I(\eta < 0))$. This loss function puts the weight $\tau - 1$ (τ) for all observations below (above) the quantile τ. Therefore, the coefficients of the τ^{th} quantile solves:

$$\begin{aligned}(\beta_l^\tau, \beta_c^\tau, \beta_k^\tau, \beta_\Phi^\tau) \quad = \quad & argmin_\beta \sum_{t=1}^{T} \sum_{i=1}^{n} \rho_\tau \Big(y_{it} - \beta_0 - \beta_l l_{it} - \beta_c c_{it} - \beta_k k_{it} \\ & -\Phi(s_{it}, n_{it}, n_{i,t-1}, c_{i,t-1}, s_{i,t-1}; \beta_\Phi) \Big)\end{aligned} \quad (1.6.2)$$

Where $\hat{\Phi}^\tau(.|\beta_\Phi^\tau)$ can be computed using the estimated β_Φ^τ. Next step is to establish the second stage to estimate the coefficient of non-transactional deposit. In contrast to the proxy model, herein I assume that the productivity at each quantile follows a first-order Markov process independent of movements in other quantiles, hence its expectation can be determined by $E[\omega_{it}|\omega_{i,t-1}]$. Having this assumption, I can set up the second stage as:

$$\begin{aligned}\beta_n^\tau = argmin_{\beta_n} \quad \sum_t \sum_i \quad & \rho_\tau \left(\zeta_{it} \right) \\ \omega_{it}^\tau \quad = \quad & \hat{\Phi}_{it}^\tau - \beta_n n_{it} \\ \zeta_{it} \quad = \quad & y_{it} - \hat{\beta}_l^\tau l_{it} - \hat{\beta}_c^\tau c_{it} - \hat{\beta}_k^\tau k_{it} - \beta_n n_{it} - \hat{E}[\omega_{it}^\tau | \omega_{i,t-1}^\tau]\end{aligned} \quad (1.6.3)$$

Here, the expectation operator is a third order polynomial of the past productivity at the quantile τ. This stage has many details that are based on the preceding assumptions. First, given β_n at each iteration, the corresponding productivity is calculated by the second line of the above equation. Next, given these values of productivity the functional expectation $\hat{E}[.|.]$ is constructed, and then, the loss function is computed at this iteration. An alternative procedure is to relax the independent assumption of the different quantiles and to allow for a cross-sectional relationship between productivities. This assumption is more realistic than my original assumption; however, it is not likely essential for addressing the size effects in scale economies. Moreover, the introduction of a cross correlation between the quantile productivities would greatly increase the computational burden because all quantile β_n^τ must be estimated at the same time. Furthermore, a panel quantile method must be used to produce a quantile operator $\hat{Q}^\tau[\omega_{it}|\{\omega_{i,t-1}^s\}_{s=0}^1]$ instead of a simple expectation. The comparison between these two assumption is beyond the scope of this paper.

1.7 Quantile Estimation Results

This section presents the results of the quantile OLS and the quantile proxy method developed in the last section using the same banking data as section 1.3. Table 1.9 shows the coefficients of a quantile regression of output on inputs in which the results are based on the assumption that the inputs are not correlated with the error term. Interestingly, this picture demonstrates that there is no evidence of significant cross sectional variation in the scale economies. In addition, the OLS coefficients reported in the last column of the table lie around the result of the 90^{th} quantile. This is not surprising when the top 10 % of bank controls over 85% of total banking activities, ordinary regression puts more weight on large banks, which may bias our conclusion for other banks.

Moreover, the OLS approach, both ordinary and quantile method, ob-

Table 1.9: Quantile Banking Production Function.

$\tau =$	0.01	0.1	0.3	0.4	0.5	0.6	0.7	0.9	0.99	0.995	OLS
labor	-0.013	-0.002	0.013	0.022	0.033	0.045	0.060	0.110	0.120	0.123	0.157
	(0.002)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.002)	(0.029)	(0.048)	(0.002)
check. dep.	0.221	0.210	0.193	0.183	0.172	0.158	0.139	0.068	0.015	0.002	0.080
	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.002)	(0.010)	(0.022)	(0.001)
Non-trans. dep.	0.682	0.659	0.645	0.635	0.620	0.598	0.565	0.380	0.085	0.058	0.352
	(0.003)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.006)	(0.017)	(0.017)	(0.003)
capital	0.119	0.141	0.165	0.179	0.200	0.228	0.268	0.485	0.822	0.850	0.457
	(0.003)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.006)	(0.022)	(0.032)	(0.003)
scale	1.009	1.007	1.015	1.020	1.025	1.029	1.033	1.044	1.042	1.033	1.046
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.006)	(0.022)	(0.002)

Data from Call Report 1993-2007, Quarterly Data, Method is Quantile Least Square Method

tains a refutable prediction for the input shares. For instance, the OLS indicates that the non-transactional deposit contributes about 60% of the small banks production, while this share is decreased for very large banks . This definition is in contrast to the picture of table 1.8 that the share of the non-transactional deposit in the total bank's liability is increasing by size or at least remains constant. Besides, table 1.8 shows that the share of capital in the production is decreasing as banks become larger. However, the growing impact of capital predicted by the OLS accords well with the theory that large banks employ capital more effectively to fulfil reserve requirements and to develop their hierarchical organizations.

In contrast to the quantile OLS, the results of the quantile proxy method are quite different from those of the previous regressions. As shown in table 1.10, the quantile proxy regression reveals that the scale economies are about flat for a substantial portion of the industry but for very large

Table 1.10: Banking Production Function (Proxy Method)

$\tau =$	0.01	0.1	0.3	0.4	0.5	0.6	0.7	0.9	0.99	0.995	QP
labor	-0.019	-0.008	0.002	0.008	0.015	0.024	0.034	0.063	0.084	0.087	0.085
	(0.002)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.015)	(0.012)	(0.002)
check. dep.	0.223	0.217	0.207	0.201	0.194	0.184	0.170	0.117	0.056	0.046	0.109
	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.004)	(0.009)	(0.005)
Non-trans. dep.	0.345	0.326	0.315	0.304	0.301	0.298	0.294	0.279	0.407	0.521	0.422
	(0.003)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.002)	(0.196)	(0.190)	(0.017)
capital	0.121	0.136	0.149	0.158	0.170	0.188	0.214	0.336	0.559	0.627	0.328
	(0.003)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.015)	(0.021)	(0.003)
scale	0.670	0.671	0.673	0.672	0.681	0.694	0.713	0.794	1.106	1.281	0.946
	(0.005)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.195)	(0.189)	(0.019)

Banking Production function using Quantile Proxy Method. The proxy variable is “Recovery Rates”, Data from Call Report 1993-2007.

banks. Middle size institutions have the largest diseconomy of scale, while the top 5% (about top 400 banks) experience somewhat an extensive economy of scale. Importantly, this result sheds light on the fact why many mergers occur among large banks: large parties involved in a consolidation benefit from both productivity improvements and scale economies. In addition, this implies that in the long run due to frequent mega-mergers, the industry will consist of a limited number of very large banks and many tiny banks, which is consistent with the trend in the last two decades. The consequence of this trend and dynamics of mergers are left for future research.

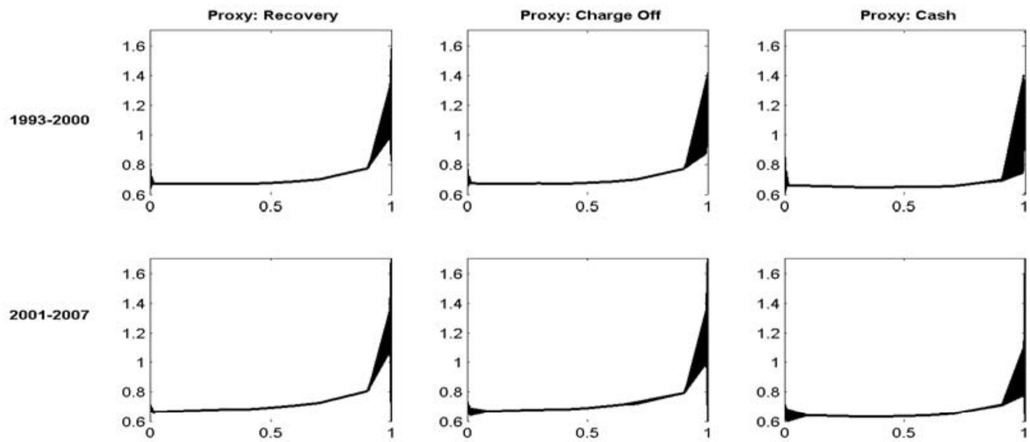
Interestingly, the shares of capital and labor in production increase with size, whereas the share of checkable deposit fades away. As I explained before, small and medium bank use their capital to issue loans, but large banks employ capital to construct their financial institutions and make a solid reputation for building up deposits. Because capital is utilized more effectively in large banks, its contribution grows as banks get larger.

Moreover, based on this estimation the Labor share in the production is rising by size. My conjecture is that because on average the size of loans are greater for big banks, and these loans are issued for large productive firms, hence the service of labor is more valuable compared to small banks. This explanation is consistent with the fact that the top banks recruit talent and pay higher wages. Furthermore, the relationship between borrowers and large banks, in contrast to small financial institutions, is vastly impersonal and formed on hard information. This allows them employ more labor in their production.

Finally, table 1.10 shows that the share of checkable deposit fades away as banks get larger. In other words, the ratio of checkable deposit on non-transactional deposit approaches to zero. Despite this structural shift, interestingly, the sum of their coefficients is constantly around 0.5, which means that these two types of deposits are well substitutable. Indeed, as argued before, non-transactional deposit contains many different types of accounts with various characteristics. The question of the role of each type account is beyond the scope of this paper and left for future research.

My finding coincides with conclusions drawn by [16]. They estimate a most preferred utility function for managers, in contrast to the conventional cost function estimation. Significantly, they document solid evidence for scale economies of very large banks. Nevertheless, their paper differs in two crucial respects with this study: the methodology and the generality of results. In

Figure 1.6: Banking Scale Economies



Scale economies for all quantiles with different proxies and for years before and after 2001

particular, they cannot find any evidence for scale economies for the pre 2001 banking data. To compare my method with theirs, I use their sample and divide observations to two groups: post and pre 2001. Next, I apply my method to the both time periods and draw the economy of scale for all quantiles at figure 1.6.

Interestingly, the estimation for both periods shows extensive scale economies for a few large banks. Although for the pre 2001 sample the standard error is larger than post 2001 result. This may stem from many financial innovations or policy deregulations in 1990's including: the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994 and the Gramm-Leach-Bliley Act of 1999. In short, the former permits acquisition and merger from other states and the latter repeals last vestiges of the Glass Steagall Act of

1933 and allows banks to engage in underwriting and selling insurance and securities. The large standard error likely prevents Hughes-Mester from come up with any scale economies for the pre 2001 sample

1.8 Conclusion

This paper estimates production functions to study scale economies and productivity in banking. Compared to the cost function method, the advantage of the production function estimation is to control for unobserved variables. Ignoring the endogeneity of unobserved variables can potentially lead to biased estimates. Another advantage is mainly data-driven. The cost function method relies on measures of prices that are generally unreliable. In contrast, the production function method uses input-level data.

I estimate production functions using two approaches. The first method assumes unobserved variables are orthogonal to input levels and employs a standard OLS regression to estimate production functions. The second approach allows for potential correlation between input levels and unobserved bank-specific productivity shocks. [29] introduce a proxy method to control for this correlation and show that OLS regression results are biased. I adopt their method in banking by introducing three new proxy equations: recovery rates of non-performing loans, charge off rates, and cash holdings of banks.

The OLS approach shows scale economies in banking. In contrast, the proxy method illustrates diseconomies of scale, which documents that OLS estimates are biased. Moreover, the proxy model rules out scale economies as a key explanation for bank mergers; however it provides evidence that merging parties benefit from the productivity growth and improvements in efficiency. The data also support the view that merging parties benefit from efficiency. After mergers, banks experience rapid asset growth and higher income per asset.

Although, the proxy model solves the simultaneity problem, it assumes that the correlation between input levels and unobserved productivity shocks is identical for different sizes. In banking, this assumption is not valid, because large banks use different technologies compared to small banks. In other words, this assumption may potentially produce biased results for very large banks. To solve this problem, I develop a quantile proxy method that can estimate a size-dependent production function. The results of the quantile proxy approach show that very large banks operate with extensive scale economies. Interestingly, this result sheds light on the fact why many mergers occur among large banks: mega-merger parties benefit from both improvements in productivity and scale economies.

Chapter 2

Business Cycle Accounting in a Small Open Economy [1]

2.1 Introduction

There has been a growing interest in applying the business cycle methodology of [7] (henceforth CKM) to study macroeconomic fluctuations in emerging economies ([37], [14], [21]). Emerging markets are of particular interest for researchers, because their business cycles are qualitatively different from those in the developed countries. In particular, in emerging economies the volatility of consumption is greater than the volatility of output and trade balance is strongly counter-cyclical. A natural framework to study fluctuations in emerging markets is a Small Open Economy.

Two important macroeconomic variables in an open economy are trade balance and current account. They are often at the center of interest in emerging markets due to the frequent phenomena of sudden stops (reversals in trade balance that often occur in recessions). The original CKM method is by construction silent about the sources of movements in trade balance, because this

[1]This chapter is an essay that was co-authored by Mohammad Hossein Rahmati and Jacek Rothert, the portion of the chapter written by the author of this dissertation is computational results

variable is sucked into the government consumption wedge.

In this paper we extend the CKM framework to a small open economy by isolating additional two wedges: (i) a trend shock wedge and (ii) a debt price wedge. We apply this methodology to shed some light on the question about the sources of fluctuations in emerging markets. Wedges are residuals of equilibrium conditions in a competitive market when data are fed into these equations. In other words, wedges that look like time-varying taxes and productivities are distorting the economy, which may have equivalent implications as frictions in a detailed economy. We look at the contribution of different wedges to the movements of different macroeconomic aggregates during the 1994 Tequila Crisis in Mexico.

In particular, our methodology allows us to evaluate relative importance of trend shocks and country risk shocks for the behavior of the business cycles in emerging markets. In two recent studies, [6] and [10] presented small open economy models encompassing the trend shock model of [1] and the country risk model of [26] and [42]—two theories of emerging markets business cycles that attracted most attention in the literature. They used Bayesian methods to estimate variances of trend shocks and country risk shocks. Both studies found that the posterior distributions of the two variances suggested that trend shocks played minor role in the fluctuations in emerging markets (Mexico in the first study and Argentina in the second), while the role of the country risk shocks was substantial. We view our work as complementary - we address the same question but with a different methodology.

First, we do the full accounting exercise. In order to do it, we use more data. In addition to the consumption, investment, output and trade balance data, we use data on hours worked and current account. We show that the discrepancy between net exports and current account provides information about the interest rate wedge. Using series on hours worked is also very informative, as this variable behaves quite differently in emerging and in developed economies. The correlation of hours worked with output is about 0.85 (s.e. 0.05) in developed countries and only 0.6 (s.e 0.12) in emerging markets (see [26]).

Second, we show how different wedges are mapped to detailed small open economies. In particular, we show that allocations in a prototype economy with interest and labor wedge, are identical to the allocations in a detailed economy with exogenous shocks to the world interest rate coupled with a working capital constraint (as in [26] or [42]). This suggests that if the dynamic behavior of an economy we study is well captured by models relying on interest rate shocks and internal frictions, than in our accounting exercise we should attribute large role to labor and interest rate wedge. The movements in these two wedges should also be highly correlated.

We also show that allocations in a small open economy with trend shocks (as in [1]) are, up to a first-order approximation, identical to allocations in a small open economy version of [2] with exogenous terms of trade shocks, if the latter are random walks. The intuition for this result is very simple. If terms of trade are random walks, then a shock to terms of trade has a permanent effect. The main difference between trend shocks and stationary productivity shocks is precisely that the former have permanent effects. To the best of our knowledge, ours is the first paper to give a formal structural interpretation of the stochastic trend. We provide evidence, that in some emerging economies terms of trade are indeed well approximated by random walks.

We are not the first to add additional wedges to the original business cycle accounting of CKM. [40] adds an asset market and a monetary policy wedge to study the relationship between output fluctuations and inflation in the US economy. [21] and [14] introduced a bond wedge to study the role of external shocks in emerging markets. Our paper is the first to introduce a wedge that is a manifestation of non-stationary shocks—a trend shock wedge.

2.2 Prototype Economy with Wedges

2.2.1 What is a wedge?

First we will give a brief explanation of what a wedge is. Consider the following first order condition describing a consumption leisure trade-off in a model with Cobb-Douglas preferences—$u(c, \ell) = c^{\eta}(1-\ell)^{1-\eta}$—and Cobb-Douglas production function—$Y = K^{\alpha}\ell^{1-\alpha}$:

$$\frac{C_t}{1-\ell_t}\frac{1-\eta}{\eta} = (1-\alpha)Y_t/\ell_t$$

In general, if we take macroeconomic data on consumption, hours worked and output, the above equation will not be satisfied. A labor wedge is a number

$\tau_{\ell,t}$ such that:

$$\frac{C_t^{data}}{1-\ell_t^{data}}\frac{1-\eta}{\eta} = (1-\tau_{\ell_t})(1-\alpha)Y_t^{data}/\ell_t^{data},$$

i.e. such that the equilibrium conditions of the model are satisfied in the data. An economic interpretation of the labor wedge is that it captures the discrepancy between the stand-in household's marginal rate of substitution between consumption and leisure, and the marginal product of labor. While the labor wedges is a reduced form concept, it may arise endogenously in a different model of a more detailed economy. For example, [7] show that the equilibrium allocations and prices in the sticky wage economy are the same as those in a prototype economy with labor wedges. Although, their prototype economy is base on a closed economy, but this conclusion can be generalized to our prototype framework, as well.

2.2.2 The prototype economy

We will now describe our prototype economy with wedges. It is a small open economy version of a standard real business cycle model of [20]. Each period, the economy experiences an exogenous shock to the state of the world s_t. The history of the shocks up to and including time t is $s^t := (s_0, s_1, ..., s_t)$.

Technology

The total output in the economy is given by:

$$Y_t(s^t) = A_t(s^t)K_t(s^{t-1})^{\alpha}\left(\Gamma_t(s^t)\ell_t(s^t)\right)^{1-\alpha}$$

where A_t is a stationary TFP shock and Γ_t is the accumulated stock of TFP, evolving according to:

$$\Gamma_t(s^t) = \gamma_t(s^t)\Gamma_{t-1}(s^{t-1}),$$

where $\log \gamma_t$ will be interpreted as a *trend shock* wedge and $\log A_t$ is the original *efficiency* wedge, as in CKM. The specification of the final output is identical to the one in [1].

Households

Stand-in household solves:

$$\max E_0 \sum_{t=0}^{\infty} \beta^t U(C_t(s^t), 1 - \ell_t(s^t))$$

s.t.

$$C_t(s^t)+(1+\tau_{x,t}(s^t))X_t(s^t) \leq (1-\tau_{\ell,t}(s^t))w_t(s^t)\ell_t(s^t)+r_t(s^t)K_t(s^{t-1})+TR_t(s^t)+ \tag{2.2.1}$$

$$+q_t(s^t)D_{t+1}(s^t) - D_t(s^{t-1}) - \frac{\psi}{2}\left(\frac{D_{t+1}(s^t)}{\Gamma_{t-1}(s^{t-1})\bar{d}} - \bar{\gamma}\right)^2 \Gamma_{t-1}(s^{t-1})\bar{d}$$

$$\lim_{t\to\infty} D_{t+1}\Pi_{i=0}^{t} q_i{}^{-1} \leq 0 \tag{2.2.2}$$

$$K_{t+1}(s^t) \leq (1-\delta)K_t(s^{t-1}) + X_t(s^t) - \frac{\phi}{2}\left(\frac{K_{t+1}(s^t)}{K_t(s^{t-1})} - \bar{\gamma}\right)^2 K_t(s^{t-1}), \tag{2.2.3}$$

where $\bar{\gamma}$ is the growth rate of the economy along the balanced growth path. The log of the price of debt—$\log q_t$—will be interpreted as the *debt price* wedge. (2.2.1) is the stand-in household's budget constraint. The last term on the RHS of (2.2.1) represents the cost of debt holding and is introduced to ensure that the law of motion for debt in a linearized version of this economy is stationary[2]. This term states that there is a cost associated with holding assets in quantities that are not equal to the long-run level. Here, the long run level of the debt position is where the normalized debt equals $\bar{d}$. In the literature, this convex cost is also called portfolio adjustment cost. In that term, $\bar{d}$ denotes a steady-state level of debt in a stationary version of the economy[3]. The remaining terms in the budget constraint are consumption expenditures C_t, purchases of investment $(1 + \tau_{x,t})X_t$, labor income $(1 - \tau_{\ell,t})w_t\ell_t$, capital income r_tK_t and lump-sum transfers from the government TR_t. The term $\tau_{x,t}$ is the *investment* wedge—any distortion that makes the relative price of investment be different from 1. The term $\tau_{\ell,t}$ is the *labor* wedge—any distortion that makes marginal rate of substitution between consumption and leisure be different from marginal product of labor. Equation (2.2.2) rules out Ponzi schemes and equation (2.2.3) governs the law of motion of capital stock. We assume quadratic adjustment costs in capital accumulation.

For now, we will assume Cobb-Douglas preferences over consumption and leisure:

$$U(C, 1 - \ell) = \frac{\left[C^{\eta}(1 - \ell)^{1-\eta}\right]^{1-\sigma}}{1 - \sigma}$$

The subsequent versions of the paper will describe how our results change if we assume quasi-linear preferences of [12].

[2]See [34] for the discussion of different ways to avoid the unit root in debt holdings.
[3]The value of $\bar{d}$ is indeterminate in theory. In our exercise, we will estimate it.

Firms

A representative, competitive firm solves a static profit maximization problem:

$$\max Y_t(s^t) - r_t(s^t)K_t(s^{t-1}) - w_t(s^t)\ell_t(s^t)$$

subject to:

$$Y_t(s^t) \leq A_t(s^t)K_t(s^{t-1})^{\alpha}\left(\Gamma_t(s^t)\ell_t(s^t)\right)^{1-\alpha}$$

Government

Government balances its period-by-period budget constraint:

$$TR_t(s^t) + G_t(s^t) = \tau_{x,t}(s^t)X_t(s^t) + \tau_{\ell,t}(s^t)w_t(s^t)\ell_t(s^t)$$

Here, we abstract from government bonds. [33] show that this model is undetermined when both private agents and government hold assets. They argue that as long as the government and the private bond have the same price, what matters is the consolidated net of all asset holdings.

Resource constraint and external sector

The economy's resource constraint is:

$$C_t(s^t) + X_t(s^t) + G_t(s^t) + NX_t(s^t) = Y_t(s^t)$$

where $\log G_t$ will be the government spending wedge. Using the government and household's budget constraint, we get that net exports are given by:

$$NX_t(s^t) = D_t(s^{t-1}) - q_t(s^t)D_{t+1}(s^t) + \frac{\psi}{2}\left(\frac{D_{t+1}(s^t)}{\Gamma_{t-1}(s^{t-1})\bar{d}} - \bar{\gamma}\right)^2 \Gamma_{t-1}(s^{t-1})\bar{d} \tag{2.2.4}$$

The last term on the RHS of (2.2.4) follows from the portfolio adjustment costs. The current account in this economy is the reduction in its outstanding foreign debt:

$$CA_t(s^t) = D_t(s^{t-1}) - D_{t+1}(s^t) \tag{2.2.5}$$

In our accounting exercise, the discrepancy between current account and net exports will help us identify the debt price wedge $\log q_t$.

2.2.3 Stationary prototype economy

Because of the trend growth, some variables are non-stationary. Similarly to [1], for each such variable X, we define:

$$\tilde{x}_t := \frac{X_t}{\Gamma_{t-1}}$$

With the above definition, we can lay out a stationary version of our prototype economy. To simplify the notation, we will omit the explicit dependence of allocations and prices on the state of the economy. Household's problem in a normalized form is now:

$$\max E_0 \sum_{t=0}^{\infty} \beta^t \Gamma_t^{\eta(1-\sigma)} \frac{\left[\tilde{c}_t^{\eta}(1-\ell_t)^{1-\eta}\right]^{1-\sigma}}{1-\sigma}$$

s.t.

$$\tilde{c}_t + (1+\tau_{x,t})\tilde{x}_t \leq (1-\tau_{\ell,t})\tilde{w}_t\ell_t + r_t\tilde{k}_t + \tilde{tr}_t + q_t\gamma_t\tilde{d}_{t+1} - \tilde{d}_t - \frac{\psi}{2}\left(\gamma_t\frac{\tilde{d}_{t+1}}{\bar{d}} - \bar{\gamma}\right)^2 \bar{d} \tag{2.2.6}$$

$$\tilde{d}_{t+1} \leq \underline{d} \tag{2.2.7}$$

$$\gamma_t \tilde{k}_{t+1} \leq (1-\delta)\tilde{k}_t + \tilde{x}_t - \frac{\phi}{2}\left(\gamma_t \frac{\tilde{k}_{t+1}}{\tilde{k}_t} - \bar{\gamma}\right)^2 \tilde{k}_t \tag{2.2.8}$$

2.2.3.1 Characterization

Equilibrium allocations are characterized by the following conditions:

$$U_{c,t}(1+\tau_{x,t})\left(1+\phi\gamma_t \frac{\tilde{k}_{t+1}}{\tilde{k}_t} - \phi\gamma\right) = \tag{2.2.9}$$

$$= \hat{\beta}_t E_t \left\{ U_{c,t+1}\left[\left(1-\delta+\frac{\phi}{2}\left[\left(\gamma_t \frac{\tilde{k}_{t+2}}{\tilde{k}_{t+1}}\right)^2 - \bar{\gamma}^2\right]\right)(1+\tau_{x,t+1}) + \alpha \frac{\tilde{y}_{t+1}}{\tilde{k}_{t+1}}\right]\right\}$$

$$\left[q_t - \psi\gamma_t \frac{\tilde{d}_{t+1}}{\bar{d}} + \psi\bar{\gamma}\right] U_{c,t} = \hat{\beta}_t E_t \left\{U_{c,t+1}\right\} \tag{2.2.10}$$

$$\frac{\tilde{c}_t}{1-\ell_t}\frac{1-\eta}{\eta} = (1-\tau_{\ell,t})(1-\alpha)\frac{\tilde{y}_t}{\ell_t} \tag{2.2.11}$$

$$\tilde{y}_t = A_t \tilde{k}_t^{\alpha}\left(\gamma_t \ell_t\right)^{1-\alpha} \tag{2.2.12}$$

$$\tilde{nx}_t = \tilde{d}_t - q_t\gamma_t\tilde{d}_t + \frac{\psi}{2}\left(\gamma_t \frac{\tilde{d}_{t+1}}{\bar{d}} - \bar{\gamma}\right)^2 \bar{d} \tag{2.2.13}$$

$$\tilde{ca}_t = \tilde{d}_t - \gamma_t \tilde{d}_{t+1} \tag{2.2.14}$$

$$\tilde{c}_t + \tilde{x}_t + \tilde{g}_t + \tilde{nx}_t = \tilde{y}_t \tag{2.2.15}$$

$$\gamma_t \tilde{k}_{t+1} = (1-\delta)\tilde{k}_t - \frac{\phi}{2}\left(\gamma_t \frac{\tilde{k}_{t+1}}{\tilde{k}_t} - \bar{\gamma}\right)^2 \tilde{k}_t \tag{2.2.16}$$

where $\hat{\beta}_t := \beta\gamma_t^{\eta(1-\sigma)-1}$. Equations (2.2.9) and (2.2.10) are the intertemporal Euler equations for capital stock and external debt respectively. Equation (2.2.11) describes the trade-off between consumption and leisure, (2.2.12) is the stationary version of production function, (2.2.13) shows the relationship between net exports and changes in external debt, (2.2.14) defines the current account, (2.2.15) is the resource constraint and (2.2.16) describes the law of motion for the capital stock.

2.2.4 Wedges

Our exogenous wedges are: (i) efficiency wedge $\log A_t$, (ii) labor wedge $\tau_{\ell,t}$, (iii) investment wedge $\tau_{x,t}$, (iv) government wedge $\log g_t$, (v) trend shock wedge $\log\gamma_t$ and (vi) interest rate wedge $\log q_t$. The first four wedges are the same as in CKM. One difference is in our interpretation of the government wedge. In our specification it is now only the government spending wedge (in CKM it also included fluctuations in net exports). The two wedges we introduce are (v) and (vi) - the trend shock wedge and the country risk wedge.

2.3 The Accounting Procedure

Similarly to CKM, we assume that the mapping between the state of the economy and the vector of wedges is one-to-one, i.e. $s_t = (\log A_t, \tau_{\ell,t}, \tau_{x,t}, \log g_t, \log\gamma_t, \log q_t)'$. In other words, the wedges in period t can be employed to uniquely uncover the state of economy or s_t. The business cycle accounting procedure consists of 3 steps. In the first step we estimate the stochastic process governing the behavior the vector of wedges $\{s_t\}$. In the second step, "knowing" the stochastic process for wedges, we identify the actual realizations of wedges (i.e. the values that the vector s_t has in each period t) - we simply solve for the values of the wedges, so that the model equilibrium conditions hold exactly in the data. In the last step we evaluate the contribution of different wedges to the fluctuations of different macroeconomic aggregates. We will now briefly summarize each of these steps, with particular focus on the identification of the trend shock wedge $\log \gamma_t$ and the debt price wedge $\log q_t$[4].

Step 1: ML estimation

The first step is to estimate the stochastic process governing the behavior of wedges. Similarly to CKM, we assume the following law of motion for the vector of wedges:

$$s_t = P_0 + P s_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma), \tag{2.3.1}$$

where s_t is a 6×1 vector of wedges, $P_0 := [\log \bar{A}, \bar{\tau}_\ell, \bar{\tau}_x, \log \bar{g}, \log \bar{\gamma}, \log \bar{q}]'$ is a 6×1 vector of steady state values of the wedges, P is a 6×6 matrix of auto-correlation coefficients and Σ is a 6×6 variance-covariance matrix of exogenous shocks. All the non-zero elements in P and in Σ as well as P_0 need to be estimated. The actual number of parameters to estimate will depend on the restrictions we impose on P and Σ. In practice, since Σ is symmetric and

[4]The details of the whole procedure can be found in [7]

positive semi-definite we will estimate the lower-triangular matrix Q satisfying $\Sigma = QQ'$.[5]

State-Space Form

We solve our prototype economy by log-linearizing[6] the decision rules around the steady state and then find the law of motion for endogenous variables via the method of undetermined coefficients. We use Uhlig's Toolkit[7] for the second step. We then use the linearized laws of motion and express the dynamic system governing the economy in the state-space form:

$$
\begin{aligned}
X_{t+1} &= AX_t + B\epsilon_{t+1} \\
Y_t &= CX_t + \eta_t
\end{aligned}
$$

where

$$
\begin{aligned}
X_t &= \left[\log \tilde{k}_t, \log \tilde{d}_t, \log \Gamma_t, \frac{NX_{t-1}}{GDP_{t-1}}, \frac{CA_{t-1}}{GDP_{t-1}}, s_t, 1\right]' \\
Y_t &= \left[\log GDP_t, \log INV_t, \log \ell_t, \log GOV_t, \Delta\frac{NX_t}{GDP_t}, \Delta\frac{CA_t}{GDP_t}\right]'.
\end{aligned}
$$

Matrices A and C depend on the parameters of our linearized laws of motion for endogenous and exogenous variables, which in turn depend on P_0 and P. Matrix B depends on the variance covariance matrix Σ. We then estimate parameters of A, B and C using maximum likelihood, where the likelihood

[5]Computationally, it is much easier to inverse a on-sided matrix than a full matrix.

[6]Except for the variables that can take on negative values. In that case we assume laws of motion are approximately linear in levels rather than in logs.

[7]Available at http://www2.wiwi.hu-berlin.de/institute/wpol/html/toolkit.htm

function is calculated using the Kalman filter (see technical appendix in [7] for details). One thing to point out is that we use change in net exports and current account rather than levels. The reason is that for some parameter values, these two variables become random walks and the likelihood is not informative. [6] encounter a similar problem and deal with it in exactly the same manner. Using first differences of net exports and current account requires that we use lagged values of these variables in our state vector X.

Step 2: Identifying wedges

Having estimated the process for wedges we need to measure the wedges in the data. Much of this procedure follows [7]. We first compute model decision rules, which are functions mapping the state of the economy to allocations. In each period t the state of the economy is given by $(s_t, \tilde{k}_t, \tilde{d}_t, \Gamma_{t-1})$. Given the estimated process for s_t, we obtain model's decision rules of the form: $C(s_t, \tilde{k}_t, \tilde{d}_t, \Gamma_{t-1})$, $\ell(s_t, \tilde{k}_t, \tilde{d}_t, \Gamma_{t-1})$, etc. The wedges are chosen so that in each period t we have $C(s_t, \tilde{k}_t, \tilde{d}_t, \Gamma_{t-1}) = C_t^{data}$, $\ell(s_t, \tilde{k}_t, \tilde{d}_t, \Gamma_{t-1}) = \ell_t^{data}$ and so on[8].

Labor wedge and government wedge are easy to measure. Government wedge is taken directly from the data on government spending. Labor wedge can be obtained from (2.2.11). Identification of the remaining four wedges is more complicated. In order to identify them we need to solve, for each period, the system of 4 equations - (2.2.9), (2.2.10), (2.2.12) and an equation that

[8]In practice, our state vector also contains the lagged values of the ratios of net exports and current account to output, because we use change in net exports and change in current account rather than levels.

links net exports with the current account. We obtain the last equation by combining (2.2.13) and (2.2.14) (intuitively, we use the difference between net exports and current account to identify the debt price wedge). We need to make some assumption regarding initial condition for Γ_0. We assume that initially the de-trended GDP is in its non-stochastic steady state. Then we may obtain the value of Γ_0 by dividing actual value from the data by the de-trended model's steady state. Then, we will know the value of $\tilde{c}_1$, $\tilde{k}_1$ and $\tilde{d}_1$[9]. Then, the four unknowns are $\log \gamma_1, \tau_{x,1}, q_1$ and $\log A_1$ which can be solved for. Knowing $\log \gamma_1$ we calculate Γ_1 and can find $\log \gamma_2, \tau_{x,2}, q_t$ and $\log A_2$, and so on. The two equations to identify the trend shock wedge are (2.2.10) and (2.2.12). The identifying moment is the response of today's consumption relative to output - trend shock raises permanent income, and so consumption response is stronger. As in [1] this is simply an exploitation of the permanent income hypothesis.

Step 3: Accounting

Our accounting procedure at the conceptual level is the same as in [7]. The output of Step 2 was the series of the estimated realizations of wedges—$\{\hat{s}_t\}_{t=1}^{T}$. In Step 3 we evaluate the contribution of each wedge to the movements in macro aggregates. We do it by feeding in one wedge at a time, or in

[9]Assuming that it is the GDP that is in its steady state value in period 1 is arbitrary. Alternatively we could assume the same thing about investment or consumption. Our results will probably differ somewhat. As a robustness check we will redo our exercise using alternative identification of initial value for Γ_0.

combinations. E.g., to evaluate the contribution of the trend shock wedge alone to fluctuations during the whole sample, for each t we set $\log A_t = \log \bar{A}$, $\tau_{\ell,t} = \bar{\tau}_{\ell}$, $\tau_{x,t} = \bar{\tau}_x$, $\log g_t = \log \bar{g}$, $\log q_t = \log \bar{q}$ and for the trend shock wedges we set $\log \gamma_t = \log \hat{\gamma}_t$, i.e. the estimated realization of the trend shocks wedge. We then feed in such modified series of s_t into the model, compute the allocations using model's decisions rules and compare them with the data.

2.4 Estimation and Accounting Results

We applied our methodology to Mexican data, covering the time period 1986-2010. We used quarterly data, which gave us 89 observations total. At this stage our results are preliminary and should be interpreted with caution. In particular, we used linear decision rules to estimate the process for wedges and to measure the actual realizations of wedges. Over the time period 1986-2010, Mexico has experienced substantial fluctuations, resembling more those of the US Great Depression rather than typical US post-war business cycle. For this reason, linear approximation might not be accurate enough - it is likely Mexico experienced shocks that took it far away from the balanced growth path. The subsequent versions of the paper will include results with decision rules calculated using second order approximation and with finite element method[10].

[10]See [35] and [24]

2.4.1 Process for Wedges

The stochastic process for wedges was assumed to be a 1-st order VAR:

$$s_t = P_0 + Ps_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma) \tag{2.4.1}$$

Because of the short time period, large overall fluctuations in Mexico and the fact that we are estimating all entries in matrices P and Σ[11], our estimates of the stochastic process for wedges are quite imprecise. In particular, the standard errors on the coefficients associated with the trend shocks wedge are quite large, so we treat our preliminary results with caution. Table 2.1 presents our estimate of matrix P and Table 2.2 presents our estimate of the lower-triangular matrix Q implying the variance-covariance matrix $\Sigma = QQ'$.

Table 2.1: Estimates of the matrix P

$\log A$	1.002	-0.134	0.000	-0.001	-0.456	0.082
	(0.190)	(0.741)	(0.163)	(0.057)	(16.563)	(2.052)
τ_ℓ	-0.017	0.500	-0.086	-0.002	7.442	-0.076
	(0.640)	(0.409)	(0.196)	(0.109)	(15.851)	(6.927)
τ_x	-0.061	0.040	0.991	-0.008	0.480	-0.413
	(0.307)	(0.602)	(0.122)	(0.095)	(12.200)	(5.211)
$\log g$	0.343	-0.665	-0.164	0.804	15.256	3.633
	(1.437)	(2.972)	(0.832)	(0.424)	(92.942)	(14.483)
$\log \gamma$	0.002	-0.008	0.001	0.001	0.781	-0.016
	(0.009)	(0.039)	(0.004)	(0.002)	(0.516)	(0.150)
$\log q$	0.003	0.020	0.001	-0.001	0.062	0.958
	(0.016)	(0.038)	(0.014)	(0.005)	(1.566)	(0.287)

Standard errors in parenthesis.

[11] We did not impose any restrictions on these two matrices except for those ensuring that P is stationary and Σ is positive semi-definite.

Table 2.2: Estimates of the matrix Q

$\log A$	0.0208	0	0	0	0	0
	(0.0054)					
τ_ℓ	0.0053	0.0147	0	0	0	0
	(0.0055)	(0.0034)				
τ_x	0.0015	-0.1147	0.0464	0	0	0
	(0.1850)	(0.2428)	(0.1076)			
$\log g$	0.0075	0.0179	0.0011	0.0536	0	0
	(0.0373)	(0.0202)	(0.1069)	(0.0122)		
$\log \gamma$	0.0001	-0.0006	0.0004	0.0001	0.0001	0
	(0.0006)	(0.0009)	(0.0006)	(0.0010)	(0.0003)	
$\log q$	-0.0001	-0.0002	-0.0004	0.0002	0.0001	0.0007
	(0.0004)	(0.0004)	(0.0012)	(0.0011)	(0.0029)	(0.0011)

Standard errors in parenthesis.

Having estimate the stochastic process governing the wedges and the expectations of agents in our prototype economy, we can measure the wedges. The two wedges we are interested in are the trend shock wedge $\log \gamma_t$ and the debt price wedge $\log q_t$. Our estimated series of the two wedges are presented in Figure 2.1. The blue line plots the estimated series for the trend shock wedge and green line plots the series for the debt price wedge. Notice the sharp decline in the trend shock wedge around the Tequila Crisis in 1994.

A somewhat surprising result is the estimate of the series for debt price wedge. We expected to notice a substantial drop around the Tequila Crisis, reflecting a sharp increase in the country risk. Our identification of the debt price relied on the discrepancy between net exports and current account. In our specification, part of that discrepancy followed from the quadratic portfolio adjustment cost. This suggest that (i) (ii) the value of the parameter ψ and

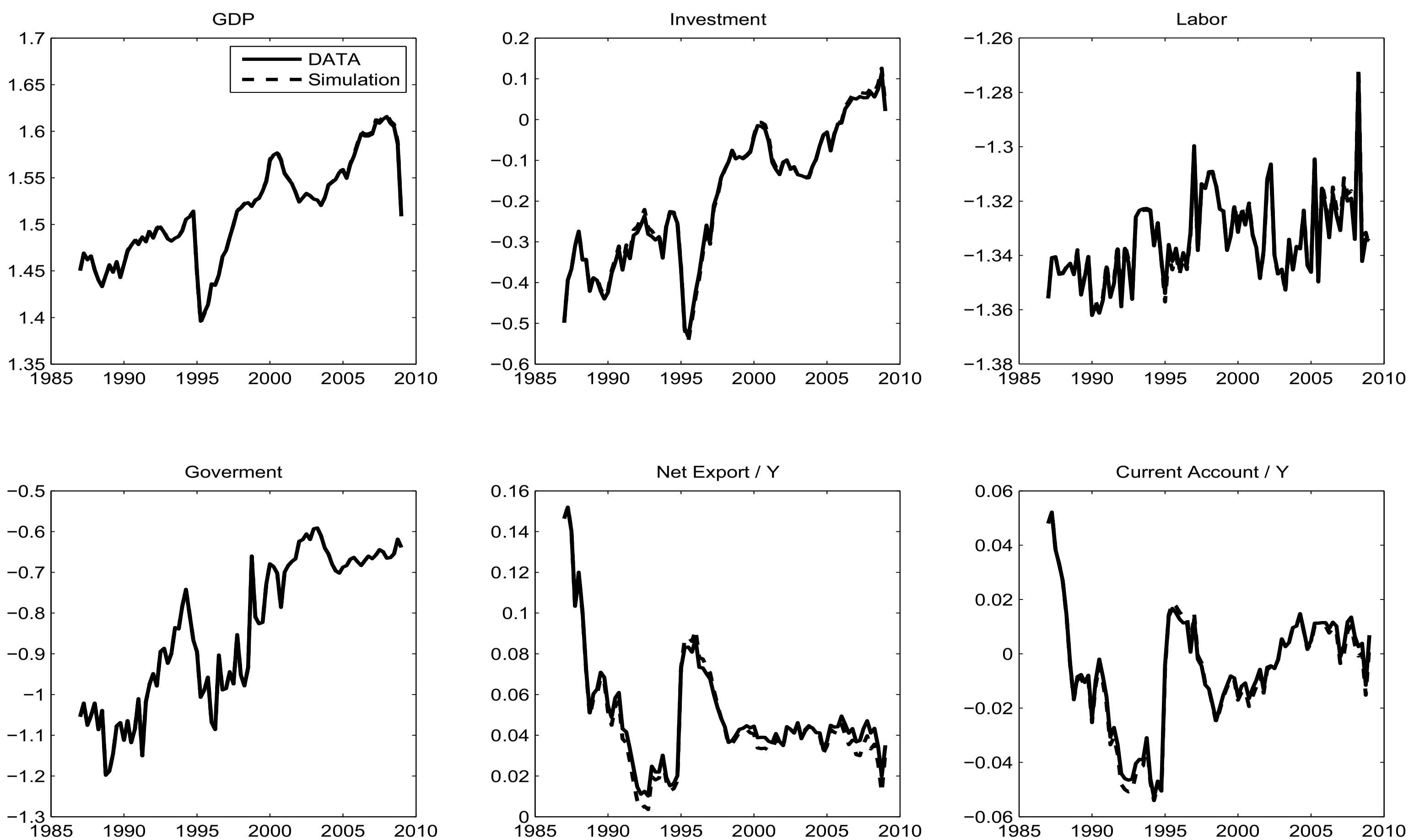
Figure 2.1: Estimated series for $\log \gamma_t$ and $\log q_t$.

(ii) the form of ensuring stationarity may affect our estimates of the debt price wedge. We have not yet evaluated the robustness of our finding to those alternative specifications.

2.4.2 Accounting for the Tequila Crisis

Before we do the accounting exercise we compare the original data with the series generated by the model where we feeded all the wedges with their values estimated in the second step of our procedure, and in which agents "know" the stochastic process for wedges is (2.4.1). Given the fact that fluctuations are substantial and that we only use 1st order approximation in both our estimation of (2.4.1) and in the calculation of the wedges, we view the almost perfect fit presented in Figure 2.2 as a huge success.

Figure 2.2: Data and simulated series with all wedges

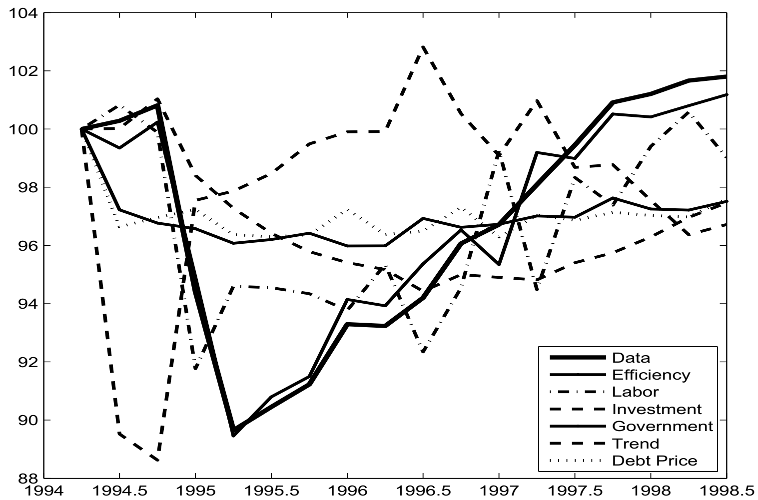


Next we look at the Tequila Crisis. In this exercise we set all the wedges but one to their steady state values. The remaining wedge is assumed to take on the "actual" value as we have previously measured. We then feed in this modified realization of the s_t process and compare the model predictions with the data. For now, we only look at three macro aggregates: GDP, net exports and current account.

Output Figure 2.3 plots the results of our benchmark accounting exercise for real output in Mexico during the Tequila crisis. The blue line plots the actual data. The other lines plot model predictions with one wedge "active" at a time. The efficiency wedge is able to account for almost all the movements in real output—the green line corresponds to the model prediction where only the measured realizations of the efficiency wedge is fed in, while all other wedges are at their steady state values. The correlation with real output is remarkable.

External sector Figures 2.4 and 2.5 plot the results of the benchmark accounting exercise for net exports and current account. Both figures show that the movements in the trend shock wedge alone can account for most of the fluctuations in net exports and the current account. Movements in other wedges (in isolation) predict an earlier sudden stop and do not generate the subsequent decline in net exports and current account in the late 1990s. These results suggest that movements in these two variables are driven primarily by non-stationary shocks (e.g. to permanent income).

Figure 2.3: Mexican GDP during the Tequila Crisis.

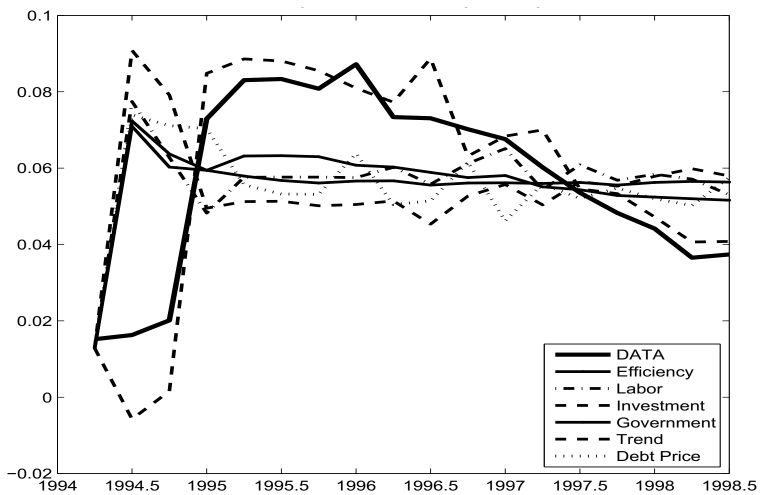


Data and model predictions for each wedge (one at a time).

2.4.3 Contribution of wedges to fluctuations during the whole sample

We also look at the contribution of different wedges to macroeconomic fluctuations during the whole sample - 1986-2010. Table 2.3 presents correlations between the data and the model simulate series with only one wedge fed in at a time. The statistics from that table confirm the results from our accounting for the Tequila Crisis. Correlations of the series for net exports and current account generated by only the trend shock movements are highly correlated with actual data. The correlation for the series of current account is particularly high—0.95. Overall, our results suggest that understanding non-stationary shocks is important for our understanding of movements in net

Figure 2.4: Mexican net exports during the Tequila Crisis.



Data and model predictions for each wedge (one at a time).

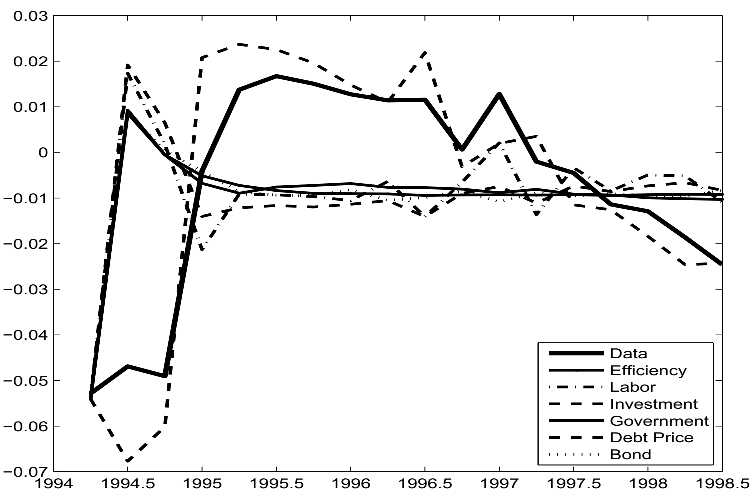
exports and current account.[12]

2.5 Equivalence Results

At the measurement level, a wedge is a discrepancy between the frictionless model's equilibrium conditions and their data counterparts (i.e. it is simply a number). It is quite meaningless, unless we can give it an economic interpretation. Typically, more than one interpretation is possible. For exam-

[12]One concern is that the underlying parameters of this experiment are stochastic variables. The reason is that the parameters A,B, and C are non-linear functions of estimated steady state levels. Therefore, an alternative test is to compute the contribution of wedges for different steady state levels and then take its weighted expectation. We leave this suggestion for future research, in particular, we aim to introduce this idea into [7]

Figure 2.5: Mexican current account during the Tequila Crisis.



Data and model predictions for each wedge (one at a time).

ple, a labor wedge is a manifestation of a friction that distorts the equality between marginal rate of substitution between consumption and leisure and the marginal product of labor. That friction may be a simple tax on labor earnings, from which government expenditures and transfers are financed. It may also be a government policy which restricts competition in the labor market as described e.g. in [8].

We introduced additional two wedges to the original CKM method.[13]

[13] [7] provide the equivalence theorems for four wedges used in their paper (i.e. efficiency wedge, labor wedge, investment wedge, and government wedge). For example, they show that a sticky wedge economy has the same allocations and prices as the prototype economy with labor wedges. Similarly, they illustrate that financial frictions in inputs may manifest themselves as efficiency wedges. Moreover, credit market frictions can be equivalent to investment wedges. All their equivalence results and subsequent proofs are also valid with

Table 2.3: Correlation of simulated series with the data

Wedge	**Y**	**NX / Y**	**CA / Y**
Efficiency	0.91	-0.08	-0.37
Labor	-0.14	-0.15	-0.21
Investment	0.03	-0.81	-0.61
Government	0.15	-0.63	-0.41
Trend	0.03	0.74	0.95
Debt price	-0.58	0.55	-0.35

All wedges set constant except one. Output is de-trended with a linear trend.

In this section we would like to give each of them a possible interpretation. First, we will show that allocations in a detailed small open economy with exogenous terms of trade shocks are, up to a first-order approximation, identical to those in the prototype economy outlined in Section 2.2, under an assumption that terms of trade are random walks. Second, we will show that a detailed economy of [26] with independent country risk and working capital constraint is equivalent to our prototype economy with debt price (almost by construction) and labor wedges.

2.5.1 A Detailed Economy with Terms of Trade Shocks

The intuition for our first equivalence result is very simple. If terms of trade are random walks, then a shock to terms of trade has a permanent effect, just like the trend shock of [1]. We will abstract from all other wedges, except

respect to the small open economy prototype. Therefore, in this section we only focused on the new wedges introduced in this paper.

for a temporary productivity shock. We will also assume away the capital adjustment and debt holding costs. This is done for simplicity and does not alter our results.

Consider the following small open economy version of [2]. The economy produces two types of goods. An intermediate good a is produced by intermediate firms (i-firms) using a Cobb-Douglas technology with inputs being capital and labor. The production function for the i-firms is:

$$Y_t^a = \tilde{A}_t K_t^\alpha \ell_t^{1-\alpha}$$

where $\tilde{A}_t$ is the stationary technology shock. The final consumption / investment good is produced using the domestic intermediate a and a foreign intermediate b. The two goods are combined according the following constant elasticity of substitution function:

$$G(a, b) = a^\omega b^{1-\omega} \tag{2.5.1}$$

The price of the imported good b is normalized to 1. The price of the country's export good a is denoted with q^a. The price of the final good is denoted with p_t. Profit maximization for the final good firms is:

$$\max p_t a_t^\omega b_t^{1-\omega} - q_t^a a_t - b_t$$

It can be shown that the price of the final good is given by:

$$p_t = (q_t^a)^\omega \left[\frac{1-\omega}{\omega} + \left(\frac{1-\omega}{\omega}\right)^{\frac{\omega}{\omega-1}} \right]$$

Terms of Trade

We assume that the economy is small, and that the price of good a exogenously fluctuates in the world market. E.g., in case of Venezuela or Russia, one can interpret a as being oil. We assume that q_t^a follows a random walk:

$$\log q_t^a = \log q_{t-1}^a + \epsilon_t^a, \quad E_{t-1}(\epsilon_t^a) = 0, \text{ all } t,$$

so that $E_t \log q_{t+1}^a = \log q_t^a$. Define:

$$\begin{aligned} \Gamma_t &= (q_t^a)^{\frac{1-\omega}{1-\alpha}} && (2.5.2) \\ A_t &= \frac{\tilde{A}_t}{p_t} && (2.5.3) \end{aligned}$$

With these definitions, we will show, that the allocations in the two economies are identical. First note that:

$$q_t^a Y_t^a = (q_t^a)^\omega \tilde{A}_t K_t^\alpha \left((q_t^a)^{\frac{1-\omega}{1-\alpha}} \ell_t \right)^{1-\alpha}$$

Hence, we will get:

$$q_t^a Y_t^a = p_t Y_t \tag{2.5.4}$$

where Y_t is GDP in economy outlined in Section 2.2.

We will also have:

$$\gamma_t = \left(\frac{q_t^a}{q_{t-1}^a} \right)^{\frac{1-\omega}{1-\alpha}}$$

We will now show the budget set and equilibrium conditions in the two economies are identical.

Stationary detailed economy

We de-trend our detailed economy in the same way as we de-trended the prototype economy, i.e. we define:

$$\hat{x}_t = \frac{\tilde{X}_t}{(q^a_{t-1})^{\frac{1-\omega}{1-\alpha}}}$$

The budget constraint for the representative household is now:

$$p_t\hat{c}_t + p_t\gamma_t\hat{k}_{t+1} \leq q^a_t Y^a_t + (1-\delta)p_t\hat{k}_t + q_t\gamma_t\hat{d}_{t+1} - \hat{d}_t \tag{2.5.5}$$

where for notational brevity we substituted $\left(\frac{q^a_t}{q^a_{t-1}}\right)^{\frac{1-\omega}{1-\alpha}} = \gamma_t$.

We will first divide both sides of (2.5.5) by p_t. Given (2.5.4), we will obtain:

$$\hat{c}_t + \gamma_t\hat{k}_{t+1} \leq A_t\hat{k}^\alpha_t(\gamma_t\ell_t)^{1-\alpha} + (1-\delta)\hat{k}_t + q_t\gamma_t\frac{\tilde{d}_{t+1}}{p_t} - \frac{\tilde{d}_t}{p_t} \tag{2.5.6}$$

Since $A_t\hat{k}^\alpha_t(\gamma_t\ell_t)^{1-\alpha} = r_t\hat{k}_t + w_t\ell_t$, and we assumed away all other wedges and capital adjustment costs, this equation looks the same as (2.2.6), except for the terms denoting debt. However, the debt in (2.2.6) is specified in units of domestic good. The debt in condition (2.5.6) above is in units of international good. Dividing it by p_t, converts it into units of today's final good, just as in (2.2.6). Hence, conditions (2.5.6) and (2.2.6) are identical (recall we assumed away portfolio adjustment costs).

The equations characterizing the equilibrium of this economy are:

$$U_{c,t} = \beta\gamma_t^{\eta(1-\sigma)-1}E_tU_{c,t+1}\left[1-\delta+\alpha\hat{k}^{\alpha-1}_{t+1}(\gamma_{t+1}\ell_{t+1})^{1-\alpha}\right] \tag{2.5.7}$$

$$q_t \frac{U_{c,t}}{p_t} = \beta \gamma_t^{\eta(1-\sigma)-1} E_t \frac{U_{c,t+1}}{p_{t+1}} \tag{2.5.8}$$

$$\frac{\hat{c}_t}{1-\ell_t} \frac{1-\eta}{\eta} = (1-\alpha) \hat{k}_t^{\alpha} \gamma_t^{1-\alpha} \ell_t^{-\alpha} \tag{2.5.9}$$

The conditions characterizing the prototype economy with trend shock wedge, assuming no other wedges and no adjustment costs, were:

$$U_{c,t} = \beta \gamma_t^{\eta(1-\sigma)-1} E_t U_{c,t+1} \left[1 - \delta + \alpha A_{t+1} \tilde{k}_{t+1}^{\alpha} (\gamma_{t+1} \ell_{t+1})^{1-\alpha} \right\} \tag{2.5.10}$$

$$q_t U_{c,t} = \hat{\beta}_t E_t \left\{ U_{c,t+1} \right\} \tag{2.5.11}$$

$$\frac{\tilde{c}_t}{1-\ell_t} \frac{1-\eta}{\eta} = (1-\alpha) \tilde{k}_t^{\alpha} \gamma_t^{1-\alpha} \ell_t^{-\alpha} \tag{2.5.12}$$

Comparing these conditions we can see they are identical, except for (2.5.8) and (2.5.11), which is why we can only talk about the equivalence in terms of dynamics of the economy. The log-linear approximation of (2.5.8) yields:

$$\hat{q}_t + \frac{U_{cc}}{U_c} \hat{c}_t + \frac{U_{c\ell}}{U_c} \hat{\ell}_t - \hat{p}_t \approx (\eta(1-\sigma)-1)\hat{g}_t + \frac{U_{cc}}{U_c} E_t \hat{c}_{t+1} + \frac{U_{c\ell}}{U_c} E_t \hat{\ell}_{t+1} - E_t \hat{p}_{t+1},$$

where $U_c := \partial U / \partial c$ and $U_{cc} = \partial U_c / \partial \log(c)$, $U_{c\ell} = \partial U_c / \partial \log(\ell)$, all of them evaluated at the steady state. Under the assumption of a random walk for q_t^a, the price of domestic consumption is also a random walk and hence $E_t \hat{p}_{t+1} = \hat{p}_t$. Hence, the log-linear approximation of (2.5.8) is:

$$\hat{q}_t + \frac{U_{cc}}{U_c} \hat{c}_t + \frac{U_{c\ell}}{U_c} \hat{\ell}_t \approx (\eta(1-\sigma)-1)\hat{g}_t + \frac{U_{cc}}{U_c} E_t \hat{c}_{t+1} + \frac{U_{c\ell}}{U_c} E_t \hat{\ell}_{t+1}$$

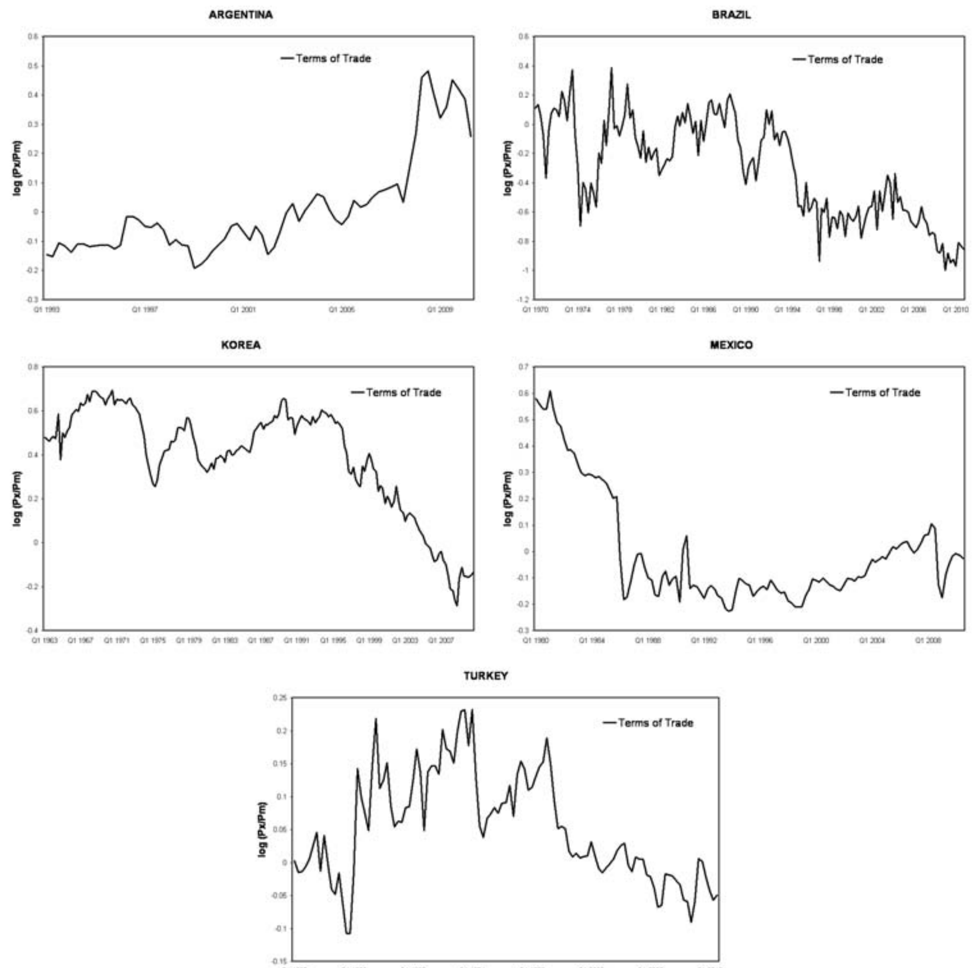
which is identical to the log-linear approximation of (2.5.11).

2.5.1.1 Are Terms of Trade Random Walks?

Our equivalence result assume that terms of trade were well approximated with a random walk process. We verified that in the data for a sample of emerging economies this is indeed a case. We collected the data on terms of trade for Argentina, Brazil, Mexico, Korea and Turkey (a sample of often studied emerging economies). Figure 2.6 in the Appendix plots the series for terms of trade for the 5 economies. For each of them performed Augmented Dickey-Fuller tests to establish whether the series contained a unit root. For none of these countries could we reject the null of a unit root. Next, to make sure the series are in fact random walks, we tested whether the series of first differences are white noise. For three out of five countries—Korea, Mexico and Turkey—we could not reject the null that the first differences of the terms of trade are white noise. Hence, given the series do contain the unit root, the terms of trade seem to be well approximated by a random walk. The results of our tests are presented in Table 2.4.

The fact that only for three countries could we make an argument for a random walk-like behavior of the terms of trade, only suggests that our interpretation of trend shocks might not work for every single country. We never hoped this would be the case. The fact that there are countries for which this can be a plausible story is very encouraging and creates an interesting avenue for further research.

Figure 2.6: Terms of Trade in Emerging Economies



2.5.2 An Economy with Working Capital Constraint

In this section we will show how an economy with interest rate shocks and working capital constraint described e.g. in [26] and [42] can be mapped to our prototype economy with debt price and labor wedge. In models with working capital friction, a fraction θ must be paid, before the production takes

Table 2.4: Terms of Trade Random Walk Test

	Unit root test		White noise tests for 1st differences			
		MacKinnon	Bartlett's		Portmanteau	
	Z-statistic	p-value	B-statistic	p-value	Q-statistic	p-value
Argentina	-1.999	0.287	1.905	0.001	65.443	0.000
Brazil	-1.633	0.466	2.246	0.000	23.682	0.000
Korea	1.877	0.999	0.833	0.492	1.680	0.794
Mexico	-0.565	0.879	0.471	0.980	2.250	0.690
Turkey	0.055	0.963	0.704	0.704	0.112	0.999

Testing for a random walk behavior of the terms of trade in emerging economies.

place. This implies firms at the beginning of time t must borrow funds to finance part of their wage bill, at interest rate R_{t-1}. The profit maximization problem for the firms is then:

$$\max A_t F(k_t, \ell_t) - w_t \ell_t - r_t k_t - (R_{t-1} - 1)\theta w_t \ell_t$$

From the firm's problem we get that the wage is given by:

$$w_t = \frac{1}{1 + \theta(R_{t-1} - 1)} A_t F_2(k_t, \ell_t)$$

and the consumption-leisure trade-off becomes:

$$\frac{-U_{\ell,t}}{U_{c,t}} = \frac{1}{1 + \theta(R_{t-1} - 1)} A_t F_2(k_t, \ell_t)$$

The above condition is equivalent to (2.2.11) if we define the labor wedge to be:

$$\tau_{\ell,t} = 1 - \frac{1}{1 + \theta(R_{t-1} - 1)}$$

With the labor wedge defined above, the detailed economy with working capital friction is equivalent to our prototype economy if, in addition we have:

$$\frac{1}{q_t} = R_t$$

which implies a tight link between the labor wedge and the debt price wedge:

$$\tau_{\ell,t} = 1 - \frac{1}{1 + \theta \left(\frac{1}{q_{t-1}} - 1 \right)}$$

2.6 Conclusions

We studied the fluctuations in a small open economy model using the business cycle accounting methodology of [7]. In doing so, we extended the original CKM method and identified fluctuations in two more wedges - debt price wedge and trend shock wedge. Our results showed that movements in the trend shock wedge alone can account for as much as 80-90% of fluctuations in net exports and current account, during both the Tequila Crisis in Mexico and during the 1980s recession in Canada.

Our results suggest that in order to understand movements in net exports and current account, we must understand the source of non-stationary shocks. We provide one interpretation of such shocks for an open economy—exogenous movements in the terms of trade, under the assumption that terms of trade are random walks. This is of course only one possible interpretation and in general we believe that our results call for a more satisfactory theory of a stochastic trend.

Chapter 3

Demand Estimation for the Iranian Automobile Industry [1]

Introduction

The Iranian auto market is supplied by two major companies, I.K.Co. (Iran Khodro Company) and Saipa-Yadak, and other relatively small producers and importers. In particular, the two largest companies account for more than 98% of the total market share, however, their products are often viewed as low quality. Interestingly, these companies sell their products with relatively high markups which is commonly observed in monopoly markets.[2] To explain these observations, this paper provides a simple theoretical framework to present the conditions under which both firms choose to produce low quality products with high prices.

Moreover, empirical estimations based on discrete choice models are used to test theoretical results by estimating marginal costs and markups and more importantly to conduct counterfactual experiments. Interestingly, the

[1] This chapter is an essay that was co-authored by Mohammad Hossein Rahmati and Seyed Reza Yousefi, the portion of the chapter written by the author of this dissertation is empirical results. An earlier version of this essay is to be published in [32]

[2] To put it concretely, markup rates of domestic cars are significantly higher than imported cars

results provide evidence for significant markups of domestic cars compared to imported cars. Also, estimated marginal costs implied by the BLP method are used in a counterfactual analysis to show that firms are reluctant to produce higher quality goods, a claim supported by the simple theoretical framework, as well. Moreover, by using these estimates of marginal costs, we show that firms may produce hight quality cars by a reduction in tariff rates. These finding are quite robust to alternative assumptions of market conduct. The critical implication for policymakers is that a reduction in the tariff can potentially increase welfare.

The estimations in this paper also investigated the market from other perspectives. Using the Multinomial Logit model, effects of changes in the attributes of people in the market were probed. The estimations suggested that an increase in average income, a decrease in average family size of the Iranian households and a decrease in the average population age will result in higher annual sales. We discuss that these estimated marginal effects of the changes in the attributes on demand could be advantageous for manufacturers. They could be beneficial for policy making strategies especially in the cases of changes in the Iranian social welfare and gradual changes in the Iranian family attributes.

The organization of the paper is as follows. Section 3.1 provides a theoretical model to explain why firms are manufacturing lower quality products. Section 3.2 explains the data used in empirical estimations. Empirical frameworks, the basic Multinomial Logit model and the BLP method, are described in Section 3.3. Section 3.4 provides three sets of empirical results. First, the results of estimations using the Multinomial Logit model are provided to explore the characteristics of the market. Second, the results of the BLP method illustrate the presence of higher markups for lower quality products in the Iranian auto industry. And third, based on the estimated parameters of the BLP model, a counterfactual analysis is performed to show that producing lower quality products is the dominant strategy for both firms in the Iranian auto market. Finaly, Section 3.5 concludes and outlines questions for future research.

3.1 Theoretical Model

This section provides a theoretical framework to study the Iranian auto market. A static game is used to obtain and to illustrate under what conditions both firms in a duopoly market choose to produce low quality products rather than highly differentiated cars. Based on the model, we conduct a simple counterfactual empirical experiment in Section 3.4.3 to support the idea that it is profitable for neither company to deviate to any other outcome or under any conduct, either collusive or oligopolistic. To build a proper model to best fit the Iranian auto market we need to make few of assumptions.

Assumption 1. The game consists of two stages. In the first period, each firm chooses to produce either a high or a low quality product, and in stage two, each firm strategically chooses the price to maximize its profit.

Assumption 2. Firms may collude in prices given their production portfolios.[3] [4]

The assumption states that after a firm chooses to produce a quality, either low or high, it is possible to collude with the other firm on pricing without breaking any laws. For simplicity, we assume that the firms choose to produce either a low or a high quality rather than a quality from a continuous state space. In addition, the assumption that firms can collude with no cost is reasonable since there is no antitrust law in Iran, i.e., no question of legality for colluding firms. Such a market leads to collusive equilibria in which firms charge high prices that approach the monopoly markups. Additionally, empirical estimates of markups in Section 3.4 verify that the companies exercise high markups in the market.

In addition to the fact that firms charge high markups, the Iranian auto market is widely known as a closed or a domestic market. In 2005, over 98% of the cars in the market were manufactured by domestic firms and less than 2% of the cars were imported to Iran by foreign producers.[5] As a result, we claim that imported cars have almost no effect on the quality choice of the major domestic producers, and for now we ignore their effects. Besides

[3] Appendix .1 provides a dynamic version of this quality-price game such that producers have no incentive to deviate from the collusion strategy

[4] An alternative assumption is that firms collude in quality and compete for prices. This framework leads to the high-low quality in the domestic production, which is not consistent with reality. Moreover, looking at the history of these firms I.K.Co established about 20 years earlier than Saipa and has manufactured about the same quality cars since its foundation.

[5] Any car assembled outside the country is called an imported car, otherwise they are called domestic. Imported automobiles are distinguished by asterisks in Table 3.6.

Table 3.1: The Quality Game

		Firm 1	
		low	high
Firm 2	low	π^l_{ll}, π^l_{ll}	π^l_{lh}, π^h_{lh}
	high	π^h_{hl}, π^l_{hl}	π^h_{hh}, π^h_{hh}

The matrix representation of the model, π^l_{hl} stands for the profit of a firm producing a low quality product while its competitor produces a high quality good, i.e., profit of the low quality producer in a high-low outcome.

these assumptions, lack of dynamic panel data prevents us from studying the dynamic strategic behaviors of the firms. Therefore, we restrict ourselves to a static framework.

As mentioned, our benchmark model consists of two periods. In period one, the firms choose the quality of the product from a list of qualities, either high or low, and pricing decisions are made in the second period. To find the subgame perfect equilibrium of the game, assumption 2 suggests that the firms choose to collude on pricing in the second period regardless of their quality choice in the first period. We denote the qualities of the higher product and the lower product by q^h and q^l, respectively, where $q^h > q^l$. Furthermore, the fixed cost of producing a high quality product is c^h which is greater than the corresponding fixed cost of producing a low quality one denoted by c^l. Our model focuses on a vertical differentiation structure where the consumers' choices vary with their taste parameters. The taste parameter of the individuals are assumed to be drawn independently from a uniform distribution with support $[\theta_l, \theta_h]$. Consumers have utility function of the form $u((p^i, q^i), \theta) = \theta q^i - p^i$, as in [41] where $i \in \{l, h\}$. Given these prices, it is straightforward to show that there exists a consumer with taste $\theta^* = \frac{p^h - p^l}{q^h - q^l}$ who is indifferent to consuming the low or high quality product. In addition, we assume that an outside option yields zero utility for consumers. It is clear that the agents with tastes lower than $\theta_* = \frac{p^l}{q^l}$ never consume any product and end up with an outside option and its zero utility. Obviously, for the existence of an equilibrium, it is necessary that both θ^* and θ_* belong to $[\theta^l, \theta^h]$. In all, we can state that:

Lemma 1. *When both high and low quality products are present in the market, given prices p^l and p^h, customers with tastes belonging to $[\theta^*, \theta^h]$ purchase the high quality product, customers with tastes $[\theta_*, \theta^*]$ purchase the low quality product and all other consumers with tastes $[\theta^l, \theta_*]$ don't buy anything and end up consuming an outside option and getting zero utility.*

Lemma 2. *When only one product quality is present in the market, i.e., both firms produce the same quality good, given quality q and price p of the good, customers with tastes $[\frac{p}{q}, \theta^h]$ purchase the product and all other consumers $[\theta^l, \frac{p}{q}]$ don't purchase anything and obtain zero utility as long as $\frac{p}{q} \geq \theta^l$.*

The model has three potential types of equilibria in the pure action space: both firms producing high quality goods, both firms producing low quality goods and the case in which one firm produces a low quality good and the other manufactures a high quality one. These three sets of strategy pairs are denoted by high-high, low-low and high-low, respectively. The normal form of the game is presented in Table 3.1 with profits of firms in the boxes for each strategy pair. In Section 3.4, we provide empirical estimations of values of the profits in Table 3.1 corresponding to each strategy-pair. The estimates suggest that the unique equilibrium is when both firms produce low quality goods in the market. Indeed, the theoretical results should be in agreement with the counterfactual experiments. Therefore, we provide conditions for the structural parameters of the model and show that for a reasonable range of parameters, both firms produce low quality goods and low-low is the only static Nash-Equilibrium of the model. To do so, we show that for all strategy pairs other than low-low, there is at least one firm with a profitable deviation from the according action/quality choice; producing a high quality product is always a strictly dominated action in our model. In other words, when firms produce different qualities, it is profitable for the high quality manufacturer to deviate and produce the lower quality good. Similarly, when both companies produce high quality goods, deviation to produce a low quality good is a profitable strategy. In all, we conclude that the only sustainable equilibrium is low-low.[6] The very first step to support the idea of the unique equilibrium is to characterize the corresponding profit of each outcome, and to obtain conditions and suitable ranges for the parameters.

First, we characterize equilibrium prices and the corresponding profits in the market when one firm produces a high quality product and the other manufactures a low quality one. As mentioned earlier, all consumers with taste $\theta \geq \theta^*$ choose the high quality product, resulting in a profit equal to

[6] To put it more concretely, the domestic firms produce low quality cars and forgo high quality markets for importers who face an extensive tariff rate

$\pi_{hl}^h = (\theta^h - \theta^*)(p^h - c^h)$ for the high quality good producer. On the other hand, all consumers belonging to $\theta^* > \theta \geq \theta_*$ choose to consume low quality goods and the profit $\pi_{hl}^l = (\theta^* - \theta_*)(p^l - c^l)$ is obtained by the low quality good producer. Based on Assumtpion 2, p^h and p^l are chosen to maximize collusive profits, that is p_{hl}^h and p_{hl}^l maximize $\pi_{hl}^l + \pi_{hl}^h$.[7] However, note that π_{hl}^h is obtained by the high quality manufacturer and π_{hl}^l is the profit of the low quality good producer.[8] According to [22], cost functions should be convex in qualities in this framework. Hence, for simplicity we assume $c^i = q^{i^2}$ where $i \in \{h, l\}$. With this assumption one can solve the optimization problem and get:

$$\begin{array}{r}\theta_{hl}^* = \frac{q^h+q^l+\theta^h}{2} \\ \theta_{hl_*} = \frac{q^l+\theta^h}{2}\end{array} \tag{3.1.1}$$

Note that, as mentioned, for the existence of equilibrium, we should have $\theta_l \leq \theta_{hl_*}$ and $\theta_{hl}^* \leq \theta_h$. Using equation 3.1.1, the latter is satisfied if we have $q^h + q^l \leq \theta^h$. This result is intuitive and comes from the consumers' optimal choice and the convexity of the cost functions when both qualities are present. The inequality implies that for any given level of lesser product quality, there is an upper limit for quality of the superior good. For example,

[7]The side payment is not allowed in Iran; however, the firms implicitly collude on geographical locations in where they open exclusive retailers. Moreover, they collude on what products they produce. For instance, both companies recently lunch the same car, called Tondar, which is supposed to be a substitute for their most popular product. By this means the firms can reduce the industry competition. In facts, the government officially prohibits any side payment in the production of Tondar and regulates the input prices extensively.

[8]We implicitly assume that each firm reaps her return from selling her cars, i.e. collusion gains are not divided

if the higher quality, q^h, is too high, then no one will buy the product and the high quality producer must exit the market. Finally, when the optimal choices of prices are plugged in the profit functions, the corresponding profits can be characterized as:

$$\begin{aligned} \pi_{hl}^h &= \tfrac{(\theta^h - q^h)q^h(\theta^h - q^h - q^l)}{4} \\ \pi_{hl}^l &= \tfrac{(\theta^h - q^l)q^h q^l}{4} \end{aligned} \tag{3.1.2}$$

Likewise, one can easily derive the profit for the two other equilibria. For example, in the case in which both firms produce low quality products, only the consumers with tastes higher than $\frac{p^l}{q^l}$ will buy the good and other consumers will not purchase anything. One can easily consider the calculated mass of consumers as the demand function, plug it into the profit function, π_{ll}^l, and maximize it over p^l to get:

$$\pi_{ll}^l = \frac{(\theta^h - q^l)^2 q^l}{8} \tag{3.1.3}$$

Similarly, when both companies produce the higher quality product the profit of each firm can be obtained as:

$$\pi_{hh}^h = \frac{(\theta^h - q^h)^2 q^h}{8} \tag{3.1.4}$$

It is worth mentioning that Assumption 2 plays a significant role in our conclusions. Notice that when both firms produce the same good, in a competitive market with no collusion, their profits will be zero. Interestingly, in our empirical estimations, the null hypothesis that the Iranian auto manufacturers are colluding is not rejected. Therefore, we claim that collusion is an acceptable assumption in such an environment with no antitrust laws.

Lemma 3. *Producing the higher qualiy product is strictly dominated by the production of the lower quality good if:*

$$\theta^h < 2q^l + q^h < \frac{2{q^h}^2 + q^h q^l - {q^l}^2 + \sqrt{2q^l {q^h}^3 - {q^h}^2 {q^l}^2}}{2q^h - q^l} \tag{3.1.5}$$

Proof. To complete the proof, after finding the profit functions by equations 3.1.2, 3.1.3, 3.1.4, one can find conditions under which both inequalities, $\pi^l_{ll} \geq \pi^h_{hl}$ and $\pi^l_{hl} \geq \pi^h_{hh}$ are satisfied. To do so, we define $\Theta_1 = \{\theta^h : \pi^l_{ll} \geq \pi^h_{hl}\}$ and $\Theta_2 = \{\theta^h : \pi^l_{hl} \geq \pi^h_{hh}\}$ that satisfy the inequalities, respectively.[9] The sets Θ_1 and Θ_2 can be represented by the following equations:

$$\begin{array}{c} \Theta_1 = \{\theta^h : |\theta^h| < \frac{2{q^h}^2 + q^h q^l - {q^l}^2 + \sqrt{2q^l {q^h}^3 - {q^h}^2 {q^l}^2}}{2q^h - q^l}\} \\ \Theta_2 = \{\theta^h : |\theta^h| < q^h + q^l + \sqrt{2q^l q^h - {q^l}^2}\} \end{array} \tag{3.1.6}$$

One can show that $\Theta_1 \in \Theta_2$ with some algebra; if θ_h is included in Θ_1, both conditions are satisfied. Using the following inequality we get the lemma's results:

$$\theta^h < 2q^l + q^h < \frac{2{q^h}^2 + q^h q^l - {q^l}^2 + \sqrt{2q^l {q^h}^3 - {q^h}^2 {q^l}^2}}{2q^h - q^l} \tag{3.1.7}$$

□

[9]Since we are looking for the symmetric equilibrium and firms are similar, the profits corresponding to the high-low and low-high outcomes will be the same.

The lemma implies that the θ_h should be bounded above (equation 3.1.5). This condition is intuitive in the sense that if the market consists of very high taste consumers, then, producing high quality product should be profitable and the market will not end up with a low-low equilibrium. Therefore, the lemma indicates that for restricted parameter values, the duopolist firms in the Iranian automobile market produce lower quality products rather than highly differentiated goods.

Empirical Method

Based on the theory section, we propose the following approach to test empirically whether the low quality production is the dominant strategy. Indeed, the real economy is more complicated than a two-goods environment as presented in this section, and it may best be described by a differentiated Bertrand competition. Therefore, the above discussion can be generalized to a more complicated model. Assume that the firm 1 is producing J_1 products and the set of its outputs is defined as $\{1, \cdots J_1\} \equiv \Im_1$. This set can be either a high quality $\Im^h$ or a low attribute $\Im^l$. In general, her total profit from producing this portfolio is $\Pi_1(P_1; P_2, \Gamma_1, \Gamma_2) = \sum_{j \in \Im_1} \pi_j(p_j; P_1, P_2, \Gamma_1, \Gamma_2)$ where P_i is the price set of the cars produced by the firm i, and Γ_i is its set of characteristics including consumer tastes and marginal costs.

Similar to the Lemma 3, to evaluate conditions under which the low quality choice is a dominant strategy, the profits must be computed for all possible quality alternatives. In the same way, it is assumed that the firms optimize their joint profit. Therefore, their profit, for instance, when firms are producing high and low are:

$$\begin{aligned}
(P_1^*, P_2^*) = argmax_{P_1,P_2} \quad & \Pi_1(P_1; P_2, \Gamma_1 = \Im^H, \Gamma_2 = \Im^L) \\
& +\Pi_2(P_2; P_1, \Gamma_1 = \Im^H, \Gamma_2 = \Im^L) \\
\Pi_{hl}^h = \Pi(P_1^*; P_2^*, \Gamma_1 = \Im^H, \Gamma_2 = \Im^L) & \\
\Pi_{hl}^l = \Pi(P_1^*; P_2^*, \Gamma_1 = \Im^L, \Gamma_2 = \Im^H) &
\end{aligned}$$

Where $\Im^H$ and $\Im^l$ are sets of high and low quality cars. Likewise, other profits that are introduced in Table 3.1 can be computed. Nevertheless, these amounts depend on the values of Γ. Therefore, before conducting any counterfactuals, section 3.4.2 estimates demand and supply parameters, including marginal costs and consumer's tastes. The discrete choice techniques are used to estimate the demand function. Estimating the marginal costs requires additional information from the manufacturers. Since this data are not available, the following $J_1 + J_2$ first order equations can be used to derive the $J_1 + J_2$ marginal costs:

$$\frac{\partial \Pi_1(P_1; P_2, \Gamma_1 = \Im^H, \Gamma_2 = \Im^L)}{\partial p_j} + \frac{\partial \Pi_2(P_2; P_1, \Gamma_1 = \Im^H, \Gamma_2 = \Im^L)}{\partial p_j} = 0$$

Importantly, estimating Γ parameters requires additional assumptions on the market conduct. The basic structure presumed in the paper is that firms jointly maximize their profits; however, for robustness check other conducts are also analyzed. For this purpose, appendix .2 replicates the counterfactual experiments when imposing different market structures on the estimation of

Γ parameters. One final remark is that we use the above equation in section 3.3.2 and derive its first order condition in equation 3.3.9.

3.2 Data

The Iranian automobile industry is primarily composed of domestic cars produced by two major manufacturers, Iran-Khodro Co. and Saipa Yadak Co.,[10] in addition to a number of limited imported cars.

Three datasets are used for the demand side which contain rich individual-level information. These micro-level observations are used to link the datasets to each other. The first dataset, gathered by the authors, contains characteristics of all cars in the market - including displacement, weight, length, width, etc. The authors have used the producers' websites to collect the necessary information on the products characteristics. The second dataset is the Consumer Expenditure Survey (CES) which is published annually by the Iranian Statistical Center. The CES dataset for 2005 contains more than 100,000 observations for the most recent 12-month period expenditures of households: specific characteristics of individuals, including family-size, education, sex, age, marital status and occupation as well as specific details about household expenditures. A valuable feature of the CES dataset is that it reports the expenditure of households on durable goods, such as cars. We have used this crucial feature to merge our datasets. Summary statistics of these two datasets

[10] Pars Khodro Company is also another major producer of domestic cars which was purchased by Saipa Yadak Co. in 2000.

Table 3.2: Summary Statistics.

	Variable	Mean	STD
Automobiles	Length (mm)	4368.29	461.09
	Width (mm)	1707.71	110.80
	Displacement (cm3)	1893.00	628.42
	Number of Cylinders	4.24	0.71
	Automatic Transmission	0.07	0.24
	*Power Steering	0.89	0.31
	*Air conditioning	0.92	0.27
	Price (000'$)	18.52	13.21
	Weight (kg)	1256.00	366.22
	*Air bag	0.32	0.47
	*ABS Braking	0.42	0.50
Number of Observations: 38			
Individuals	Sex	0.28	0.45
	Family Size	3.86	1.94
	Age	43.42	13.49
	Education	29.45	22.43
	Income (000'$)	290.00	310.00
Number of Observations: 30438			

The variables marked by asterisks are dummy variables. Sex is equal to one for female and zero for male.

are available in Table 3.2.

The third dataset is provided by the Marketing Management in I.K.Co. (MMI). This dataset contains 30,438 observations of car owners in Tehran. Data on owners' gender, family size, and occupation, along with their car models are gathered at gas stations. The dataset contains observations of all types of cars in the market including samples from imported and domestic cars in 2005.

Table 3.3: The Production and Market Share

Company	Car	Real Data Sales	Real Data Market Share	MMI Sales	MMI Market Share
I.K.Co.	Peykan	90,776	10.86	2636	9.9
	Samand	69,223	8.28	2189	8.2
	RD	73,146	8.75	2434	9.1
	GLX	119,759	14.33	3342	12.5
	Peugeot 206	75,775	9.07	2800	10.5
	Pars	37,713	4.51	1231	4.6
	Vanet	44,169	5.28	647	2.4
SAIPA (Including PARS)	Pride	291,013	34.81	10566	39.3
	Xantia	13,008	1.56	511	1.9
	Rio	2,939	0.35	13	0.05
	Caravan	765	0.09	39	0.1
	Nissan	1,858	0.23	174	0.7
	Maxima	2,359	0.28	163	0.6

Production and market share of two big companies in 2005

Table 3.3 shows real market shares of a number of cars in the country[11]. However, calculated market shares based on the MMI (Table 3.3), which were gathered in Tehran, are close to the real market shares (Table 3.3). Therefore, we have used the market shares in the MMI dataset as an approximation of the real market shares of all cars in the market. The MMI contains 38 car models, 17 of which are imported; the other 21 are produced by the duopolist companies. Prices range from 5.5 thousand dollars for Vanet to 43 thousand dollars for Merc. Chairman. The data on prices are reported by Industrial Development and Renovation Organization of Iran (IDRO).

We used their common factors to link the datasets. We chose the MMI

[11]The shares were reported by the manufacturers

as the base dataset and applied serious restrictions in merging them. To insert new attributes for individuals in the MMI, for example household incomes, we conducted a two-step procedure. First we found the distribution of income conditional on family size, occupation and type of purchased car. Then, we performed a random draw from the obtained conditional distribution. For example, to do a random draw for a buyer who is a teacher, has a family with three members and has purchased a Toyota Camry, we used CES to construct the conditional distribution of income for teachers (as occupation) who own a Toyota Camry (as type of car) and with family size equal to three. Such a method enabled us to generate a more accurate final dataset than using an unconditional random draw from aggregate distributions of the attributes.

The next section provides more insight into the data and the choices of individuals with respect to their characteristics. A Multinomial Logit model is employed to see the effects of population attributes on choices in the Iranian automobile market.

3.3 Empirical Models

Three empirical exercises are performed in this section. The Mutinomial Logit estimation shows the role of the attributes of people in their choices in the Iranian auto market. The BLP methodology was performed to estimate markups which showed that domestic firms are exercising much higher markups on their products relative to imported car manufacturers. Also, the marginal cost of each automobile was estimated in this experiment in order to perform another counterfactual analysis. This analysis suggests that it is profitable for neither firm to produce higher quality products, as stated in Lemma 3

The methods used in this paper, particularly discrete choice models, are common in the applied Industrial Organization literature. In discrete choice models, individuals choose from a set of mutually exclusive options to gain the highest possible utility. In these models, the utilities of alternatives for individuals depend both on the consumers' attributes and the characteristics of the alternatives. However, utilities can not be measured directly when some of the attributes that influence the utilities are unobservable. In addition, demographics are used as a proxy for consumer heterogeneity in such models. The Multinomial Logit model, as in [13], the Nested Logit model, as in [11], and the Random Coefficients model used by [5] are among the most common models utilized in the literature.

As previously mentioned, the Multinomial Logit model is frequently used to estimate random utility maximization problems. Under the assumptions of this model, the probabilities of choosing different alternatives are computed as functions of the attributes of the customers. Various authors have applied the Multinomial Logit model to their papers. For example, [13] apply the Multinomial Logit model in the coffee market and study the repeat purchasing behavior of the shoppers. They allow for an additional marginal utility for future consumption of the already-consumed products to explain the positive correlation among sales patterns. In an earlier application, [9] used the model in the context of transportation planning to explain the decision for alternative transportation modes (car, bus or etc.). Structurally, the Multinomial Logit model implies that substitution patterns are proportional when the choice probabilities of alternatives depend only on their own characteristics and are independent of the characteristics of other present choices. As a result, the Multinomial Logit model fails to give a reasonable explanation for substitution patterns between different choices in models where utilities are described as functions of the attributes of the choices or the consumers. However, this model provides a framework to investigate some counterfactual analyses such as the effects of population attributes on demands.

To allow for more realistic substitution patterns between alternatives with acceptable cross-price elasticities of demand, [5] and [4] introduced more flexible characteristic-based utilities.[12] In particular, market-level demand functions are obtained from aggregating over customer-level demand functions, where utility functions relate market equilibrium prices to the market shares of goods. In their framework, the interaction between supply and demand on the consumer level determines the equilibrium prices and develops new methods to estimate costs as well as the demand parameters. The method has been developed and implemented by many authors for other differentiated markets. In an interesting application, [27] and [28] used this method for the US ready-to-eat cereal market and conducted experiments to evaluate mergers

[12]In the simple Logit framework, all cross price elasticities with respect to the price of the good i are equal to each other.

in this market. Similarly, [31] considered extra moments derived from micro data to identify more precise parameters for the US auto industry.

3.3.1 Multinomial Logit

The Multinomial Logit model is used to investigate the effects of population characteristics on demand structure in the Iranian auto market. Utility from purchasing a car consists of two observable and unobservable terms. A part of utility from purchasing a car which is observable to an economist depends on the characteristics of individuals, X_i (including income, age, education, family-size and sex of the buyer in our model). For example, people with different family-sizes are more likely to pick up different cars. The general form of the utility of a person with characteristics (X_i) from choosing alternative j is expressed as:

$$U_{ij} = X_i \beta_j + \epsilon_{ij} \tag{3.3.1}$$

where the disturbance ϵ_{ij} is assumed to be of a type 1 extreme value distribution. The assumption used in the estimation is to consider mean exogenous utilities for goods and to allow utilities to vary depending on individual characteristics. These characteristics act as proxies to illustrate the substitution patterns between different cars. The mean utilities of cars are estimated through the first entry of β_j where the corresponding variable is equal to one in the X_i vector. [5] computes the probability of choosing alternative j among J available alternatives by integrating over all possible unobservable errors to calculate the probability of having utility from purchasing good j exceed utilities of all other goods:

$$prob(\text{purchase good } j|X_i) = \frac{e^{X_i\beta_j}}{\sum_{k\in J} e^{X_i\beta_k}} \tag{3.3.2}$$

Having access to individual-level data we have information on people who chose outside option, i.e., they did not purchase any of the products, at all. Letting the utility of consumption from outside option be zero, we have resolved the identification issue by comparing utilities from different alternatives with respect to the zero utility of our base product, the outside option. Estimates from the Iranian auto industry are used to illustrate the marginal effects of age and income on demand. The effects of income are of interest to us since the average Iranian real GDP is dependent on exported oil prices, resulting in a volatile per capita real income. Marginal effects of age are also considered in this simulation to investigate the effects of changes in average age of the population.

3.3.2 BLP(Berry, Levinsohn and Pakes)

A simple case of [4] is to consider the following model for utility to denote the different preferences for the characteristics of products:

$$U(\zeta_i, p_j, x_j, \xi_j; \theta) = x_j\bar{\beta} - \alpha p_j + \xi_j + \epsilon_{ij} \tag{3.3.3}$$

where the utility is composed of the error term, ϵ_{ij}, and the mean utility of purchasing a car defined as:

$$\delta_j = x_j\bar{\beta} - \alpha p_j + \xi_j \tag{3.3.4}$$

Using the assumption that error terms have i.i.d. extreme value distributions, they express market shares in terms of the mean utilities by integrating over all possible errors:

$$s_j = \frac{e^{\delta_j}}{\sum_{k \in J} e^{\delta_k}} \tag{3.3.5}$$

Using the calculated market shares and knowing that the sum of all shares should be equal to one, they back out the mean utilities from market shares as

$$\delta_j = Ln(s_j) - Ln(s_0) \tag{3.3.6}$$

where Ln is the natural logarithm function. Hence, one might estimate the following equation to obtain the elasticities of prices and characteristics of the products:

$$Ln(s_j) - Ln(s_0) = x_j\bar{\beta} - \alpha p_j + \xi_j \tag{3.3.7}$$

Furthermore, based on assuming Bertrand-Nash equilibrium in the market, they assume that the profit functions of the firms are of the form:

$$\Pi_f = \sum_{j \in \Im_f} (p_j - mc_j) M s_j(p, x, \xi; \theta) \tag{3.3.8}$$

where mc_j is the marginal cost of producing good j, M is the total number of auto purchases in the market and $\Im_f$ is the set of cars produced by firm f. Given these assumptions, any product of firm f must have a price p_j which satisfies the first order condition:

$$s_j(p, x, \xi; \theta) + \sum_{r \in \Im_j} (p_r - mc_r) \frac{\partial s_j(p, x, \xi; \theta)}{\partial p_j} = 0 \tag{3.3.9}$$

The J first order conditions are used to obtain the markups calculated by $\frac{(p_j - mc_j)}{p_j}$. To do so, a new J by J matrix is defined where its (j, r) element is given by:

$$\Delta_{jr} = \begin{cases} \frac{-\partial s_r}{\partial p_j} & \text{if } r \text{ and } j \text{ are produced by the same firm} \\ 0 & \text{otherwise} \end{cases} \tag{3.3.10}$$

to solve and obtain:

$$p = mc + \Delta(p, x, \xi; \theta)^{-1} s(P, x, \xi; \theta) \tag{3.3.11}$$

The last equation can be used to investigate the effects of any perturbations in the market structure on equilibrium prices using the estimated costs. Section 3.4.2 uses this equation to estimate marginal costs of the automobiles given market prices. In this model, the price elasticities can be calculated analytically: $\eta_{jr} = \frac{\partial s_j p_k}{\partial p_k s_j} = -\alpha p_j(1 - s_j)$ if $j = k$, otherwise $= \alpha p_k s_k$. Therefore, using these estimates of elasticities and assuming a market structure for competition, Δ_{jr} can be calculated from eqution 3.3.10. Given these estimates of Δ_{jr}, marginal costs can be derived by solving equation 3.3.11. Notice that the estimated costs depend on the assumption made for the market conduct. As a result, for the robustness check, it is necessary to estimate the marginal costs with different market structures.

The same equation is used in Section 3.4.3 to conduct a counterfactual analysis in which market structure is altered in an exercise to allow for high-low and high-high outcomes. In these experiments, the estimated marginal costs are fixed and the equation 3.3.11 is solved to derive the optimal pricing strategies for firms. In the analysis, equilibrium prices and profits are predicted using the new hypothetical market structure and the estimated costs from Section 3.4.2. They indicate that the only sustainable equilibrium is the production of low quality products by both firms.

3.4 Empirical Results

3.4.1 Multinomial Logit

The multinomial logit model described in Section 3.3 is used to illustrate the role of the attributes of consumers on utilities and demands for cars. The attributes of consumers used in the estimation are sex, family-size, age, education status and income of individuals. Estimation results provided in Table 10 are statistically significant - according to the standard errors in parentheses - implying interesting facts about the Iranian market. For example, the coefficients on the sex of consumers (a dummy variable taking the value of one for female and zero for male) suggest that females obtain higher utility from all models of Peugeot 206, Renault, Matiz, and Gol, and derive less utility from Nissan-Pickup and Sinad. This pattern of taste is widely understood in the Iranian auto market. Family-size has bigger coefficients for bigger cars and smaller coefficients for smaller cars, indicating a remarkable correlation between car size and family-size.

To interpret the coefficients of Table 10, it is instructive to examine results of two distinctive cars. For example, consider Peugeot 206, a relatively expensive supermini car targeted small families and women, compares to Peykan, a cheap economic car designed based on Hillman Hunter and are produced since 1967 by I.k.Co. Interestingly, despite its high sales, the attributes and features of Peykan have remained unchanged for all these years. First, it is obvious that the sex coefficient is significantly smaller for Peykan meaning that women prefer 206 to a large extent. Moreover, the estimates indicate that large families dislike 206, and rich households favor 206 over Peykan. Finally, the differences between education and age coefficients are not economically meaningful.

Furthermore, in Table 3.4, we have calculated the marginal effects of each attribute on the probability of choosing outside option (not purchasing any new car). Since the probability of choosing a car is proportional to its market share, marginal effects can be appropriate measures of the changes of

Table 3.4: Marginal effects for the Multinomial Logit Model

	$\frac{\partial y}{\partial x}\%$	Std. Err.	t-test	p-value
Sex	6.91	0.05	132.24	0.00
Family size	0.51	0.01	48.34	0.00
Age	0.04	0.00	34.04	0.00
Education	-0.04	0.00	-50.56	0.00
Income	-0.08	0.43	-1.95	0.05

$\frac{\partial y}{\partial x}$ is for discrete change of dummy variable from 0 to 1.

market shares due to changes in attributes of the market consumers.

The simulation results suggest that a marginal one unit increase in average family-size, a marginal one year increase in average age and a marginal one unit increase in average income - equivalent to one hundred dollars in our simulation - will result in 0.51%, 0.04% and -0.0008% change respectively in the probability that an average consumer chooses the outside option. By average consumer, we mean a consumer with characteristics equal to the average of the characteristics of all people in the dataset. According to the CES data, almost 6% of households buy new cars every year and 94% of households are people who don't buy any new cars, whom we denote as the "outside option buyers". Taking into consideration that more than 1,100,000 cars are produced in a recent year, the mentioned changes in consumer attributes may result in approximately -83000, -6700 and 130 changes respectively in the number of total annual car sales. This might be of interest from the policy making perspective for manufacturers who would like to have estimates of demand changes due to the changes in population demographics. As an example, the baby-boom phenomenon in Iran which resulted in a younger average population trend in 00's can be an explanation of the decreasing market share of "outside option" in recent years. Another interesting feature of these estimates is the demand shift under a change in average income of people. In other words, volatility of the Iranian real per capita GDP may be a proxy of fluctuation in demand over different periods which is captured by the estimates.

3.4.2 The BLP Method

The estimates from equation 3.3.7 are presented in Table 3.5 where all utility parameters have reasonable signs: negative sign for price, positive sign for safety, and negative sign for weight. These estimates illustrate that the domestic cars show an average of 4.19 for the own-price elasticity. Interestingly, the same measure exceeds 8.76 for the foreign cars. Furthermore, estimates of marginal costs for all cars were computed from equation 3.3.11 in order to obtain markups for all cars in the market. The estimates of marginal costs and markups are summarized in Table 3.6.

Comparing markups for different cars, we observe that the markups are significantly higher for domestic cars, which are distinguished by asterisks in the Table 3.6. This is an indication that domestic producers are enjoying high markups which the authors believe are mostly due to the tariff policy by the Iranian government on auto-products. In 2005, the tariff policy by the Iranian government levied 70% tariff on intermediary goods which the domestic manufacturers used in their production lines. On the other hand,

Table 3.5: The BLP Estimation

	Coef.	Std. Err.	t-test	p-value
Price (000'$)	-0.449	0.190	-2.360	0.018
Length (mm)	-0.009	1.267	-0.010	0.994
Width (mm)	-6.130	5.210	-1.180	0.239
Displacement (cm^3)	0.013	0.006	2.150	0.032
Number of Cylinders	-3.412	1.995	-1.710	0.087
Weight (kg)	-0.008	0.004	-2.140	0.033
Safety	3.610	1.866	1.930	0.053
Amenities	1.343	1.520	0.880	0.377
Constant	7.931	11.135	0.710	0.476

Note: Number of observation is equal to 38 and Wald test is $\chi^2(8) = 16.77$. The weighted average own-price elasticity for the imported cars is -8.77 (-3.44, -24.50) and for the domestic cars is -4.19 (-2.24, -16.78)

tariff on the prices of imported automobiles have consistently been 100%. In order to make the market more competitive, such that consumer surplus and total surplus both increase, the government may apply a tariff policy to reduce the market powers of domestic producers. Therefore, a decrease in tariffs for imported automobiles is a major aspect that needs further research in order to make the Iranian auto market more competitive and to achieve higher social welfare.

3.4.3 Counterfactual Experiment

Low Quality Choices

This section provides an empirical counterfactual test to show that it is not profitable for either firms to produce higher quality products in the Iranian automobile market. For this purpose, we needed to estimate the profits of firms

Table 3.6: Marginal costs and Markups for Automobiles

Car	Markup(%)	Marginal Cost($000's)	Car	Markup(%)	Marginal Cost($000's)
Peykan	37.40	3.82	Mazda*	10.02	19.80
Vanet	41.47	3.22	Pajero*	4.08	51.80
RD	33.06	4.62	Renault	44.59	2.74
GLI	22.15	8.02	Maxima	5.97	34.80
206 V2	21.42	8.37	Pickup	9.01	22.30
GLX	18.62	9.97	Vanet Mazda*	21.00	8.30
Samand	20.46	8.87	Roniz*	7.23	28.30
Samand LX	17.41	10.82	Proton*	16.20	11.40
206 v3	19.67	9.32	Sinad*	23.25	7.28
206 v5	18.40	10.12	Verna*	16.11	11.48
206 v6	16.90	11.22	Gol*	16.82	10.90
Pars	14.72	13.22	Musso*	5.21	40.12
ELX	13.04	15.22	Merc. Chairman*	4.90	42.80
Nasim	33.91	4.40	Corolla*	8.16	24.80
Saba	32.21	4.75	Camry*	5.34	39.10
Xantia	10.09	20.10	Avante*	9.18	21.80
Rio	19.78	9.15	Prado*	4.41	47.80
Caravan	18.04	10.25	Cielo*	18.53	9.70
Matiz*	29.01	5.40	Patrol*	14.70	12.80

Imported cars are marked with asterisks. Also, 206 stands for Peugeot 206 in the table.

for different strategy-pairs, i.e., the profits in Table 3.1. Also, we ranked the brands of automobiles with respect to their qualities. To do so, we used the coefficients obtained in the BLP estimation to rank the automobiles based on the weighted importance of their characteristics, $X_j\beta_j$. Then, we labeled the 19 cars with higher values of the estimated term (sum of the multiplication of their characteristics by the corresponding coefficients) as the high quality products. The remaining 19 automobiles with lower values were designated as the low quality products. Given that there is no antitrust law in the Iranian market, we assumed that it is a dominant strategy for both firms to collude in pricing. Then, we estimated the profits under different choices of qualities as follows.

First, we assumed that both firms produce high quality products. Given the estimated marginal costs of the automobiles and the market structure where only the 19 higher quality products are being manufactured, we used the first order condition matrix presented in the BLP model (equation 3.3.11).[13] This allowed us to calculate the set of equilibrium prices which gives the highest possible profit.[14] The profits are predicted by equation 3.3.8 given prices.

[13]This experiment is based on the marginal costs estimated in Section 3.4.2. One concern is that the collusion assumption made to estimate the marginal costs might misrepresent the real structure of the marker. To address this issue and to check the robustness of the results, we employ the same procedure proposed here but with different assumptions on the market structure. Appendix .2 describes these robustness checks.

[14]Notice that equation 3.3.11 is a non-linear function of prices because a new price equilibrium would change the market shares as well as the markups. It requires to solve for the all prices at the same time. This complex non-linear optimization problem is not solvable when the number of goods/markets are considerably large or if the utility has random coefficient components

Table 3.7: Profit in Quality Game

		Firm 1 low	Firm 1 high
Firm 2	low	(28.20,28.20)	(55.67,15.17)
	high	(15.17,55.67)	(7.97,7.97)

Estimated counterfactual profits in million Dollars for the Iranain auto industry. All profits are calculated by the assumption that firms are colluding.

It suggests that in such a market each firm would obtain a profit equal to 7.6 million Dollars in 2005. Second, we considered the case when a company produces higher quality products and the other produces the lower quality automobiles. Using the same profit function and estimating the equilibrium prices under the new market structure, the counterfactual analysis indicated that the higher quality producer would obtain 53.3 million Dollars and the other lower quality producer would get 14.5 million Dollars in equilibrium. Similarly, we estimated equilibrium profits for the case when both firms produce the 19 lower quality cars and predicted 27.0 million Dollars for each. Considering all the projected profits under different strategies in Table 3.7, we can see that the only sustainable Nash equilibrium of the game is when both firms produce lower quality products and the higher quality automobiles' production is a strictly dominated strategy.

Tariff Reduction

Another interesting experiment is to evaluate the impact of a tariff reduction on the market equilibrium. As stated before, the imported cars constitute about half of the brands in the market, while their sales amount only 2% of the market. They are not essentially luxury cars and include also economic cars like Camry and Mazda. Moreover, the car manufacturing is recognized as a protected industry, and thus, the Iranian government imposes a tariff rate of over 100% on importing cars. This barrier to trade appears as an additional cost to production. Therefore, it may dissuade importers from supplying in the market. In the absence of high quality cars, the low-quality domestic products entice many high-taste consumers. Consequently, there is no competition to force the domestic firms to improve the quality. As a result, there is some debate whether high tariff rates contribute to low-quality domestic products.

In order to evaluate the impact of the high tariff rate, we conduct two experiments when the marginal cost of the imported cars reduced by 20% and 30%. This doesn't necessarily imply that the tariff rates are cut by the same amount. Indeed, these measures can be reached by somewhat extensive reduction in tariffs. Finally, similar to the previous section all equilibrium outcomes are computed assuming the firms colluding on their pricing policy.[15]

Table 3.8 shows the result of a game when the marginal costs reduced by 20% and 30%. Interestingly, the left panel illustrates that a reduction

[15] An identical robustness check as Appendix .2 indicates that this result remains unchanged when estimating the marginal costs with a different set of assumptions for market conduct

Table 3.8: Profit for Counterfactual Quality Game

	20% cut Low	**20% cut** Hight	**30% cut** Low	**30% cut** High
Low	(29.35,29.35)	(56.92,35.91)	(32.00,32.00)	(44.83,19.52)
High	(35.91,56.92)	(18.87,18.87)	(19.52,44.83)	(78.16,78.16)

The counterfactural experiments with 20% and 30% cut on the marginal costs of the imported cars.

of 20% changes the equilibrium outcome from the low-low product to the high-low market. Moreover, a further reduction in the marginal costs of the imported cars leads to the high-high equilibrium. Importantly, we do not claim that by this change in the tariff rates, the domestic companies certainly produce high quality cars. In fact, we cannot answer this question because the adoption of a new technology is beyond the scope of our model and remains as an important topic for future research. What we can assert is that with the current technology a reduction in the tariff rates promotes either the more imports of foreign cars or a domestic production of high quality cars or both, such that the market is no longer supplied by low quality products.

3.5 Conclusion

The Iranian automobile market is an example of oligopolistic differentiated products market with two major domestic manufacturers and a number of importing firms. This paper considers the Iranian auto market as a case study and provides a framework to explain the structure of the market. The theoretical model shows how firms may choose to produce lower quality products rather than highly differentiated qualities. Counterfactual empirical analyses confirmed the theoretical hypothesis that both domestic firms preferred to manufacture lower quality products by comparing the projected profits under different market structures. Furthermore, empirical estimations based on BLP (1995) suggested that domestic producers are enjoying higher markups relative to the importing firms. The pattern of estimated markups also showed evidence that they are considerably higher for lower quality goods and imply that domestic companies collude in pricing their products. The estimation procedure was performed in two steps, providing precise estimations of marginal costs, and recovering the markups based on estimated costs in the second stage.

Further interesting extensions - some of which are being currently investigated by the authors - can be done by future researchers. More thorough analyses on markups and more precise estimations can be done by using panel data and taking the intertemporal production and pricing into account. In particular, future research may focus on dynamic aspects of the automobile market to more accurately separate the quality and pricing decisions of the producers. Merger analysis can be done to analyze the effects of the merger between different companies. Furthermore, using the market structure and the estimates of markups, it is interesting to analyze the profitability of the production of a common product by companies. Such an example have existed in recent years where the two largest companies(I.K.Co. and Saipa) have collaboratively started to produce a new automobile, Tondar L90.

Appendices

.1 Dynamic Theory Model

In this appendix, we propose a dynamic version of the quality-price competition similar to what is described in Section 3.1. The goal of this appendix is to relax assumption 2 and to study under what conditions firms tend to collude in pricing. To explore this issue, we assume that in period zero firms decide on what quality to produce. After the firms determine their qualities, in all following periods $t = 1, 2, \cdots$ they playing a Bertrand pricing game. Firms discount future profits at the rate δ. In particular, we study the trigger strategy equilibria, in which firms play the duopoly competition after an occurrence of any deviation. More formally, if $\sigma^{collusion}$ stands for the strategy associated with collusion, the trigger strategy is defined as:

$$\sigma_t = \begin{cases} \sigma^{collusion} & \text{if} \quad \sigma_{t-1} = \sigma^{collusion} \\ \sigma^{competition} & \text{if} \quad \sigma_{t-1} \neq \sigma^{collusion} \end{cases}$$

The Folk Theorem of repeated games establishes that there is a minimum discount rate δ^* such that for all $\delta > \delta^*$ the trigger strategy deters firms from any deviation. To find the δ^*, it requires to employ the one-shot deviation principle. In fact, three equilibria can be recognized when firms have decided to produce low-low, high-high, and low-high qualities. In case of any deviation, When firms producing same qualities they achieve zero profits because of the Bertrand competition on prices. Therefore, one can simply use the one-shot deviation principle and reach to the following results:

$$\delta^*_{ll} = \delta^*_{hh} = \frac{1}{2}$$

Moreover, when qualities are high-low, since the profit of collusion is greater than the competition, and the static profit of deviation exceeds the profit of collusion, by the Folk Theorem there exist a $\delta^* \in (0,1)$ such that for all $\delta \in (\delta^*, 1)$ deviation from collusion is not profitable. It is straightforward to show that $\delta^* = \frac{\pi^{deviation} - \pi^{collusion}}{\pi^{deviation} - \pi^{competition}}$. In sum, if both firms are enough patient then the trigger strategy can support the collusion equilibrium similar to what considered in assumption 2.

.2 Counterfactuals With Other Market Structure

In Section 3.4, the underlying assumption of the counterfactual experiment is that the two domestic firms in the Iranian auto market collude in their pricing, when the marginal costs are estimated. However, one might argue that this assumption can potentially bias the results. Therefore, in this appendix the same procedure is employed with various market structures when estimating the marginal costs. After estimating new marginal costs, the same experiment is run to obtain prices and profits for each game.

The results of this exercise is shown in Table 9. It is important to stress that the market conduct for each table is the assumption used to compute the marginal costs. Notwithstanding, the profits reported in each equilibrium is the outcome of an experiment when firms are colluding. These tables highlight that the assumption made for the market structure to compute the marginal costs has negligible effects on our conclusions. In other words, in all market structures presented in Table 9 firms prefer to produce low quality cars.

Table 9: Profit in Quality Game (Other Market Structure)

Doupoly Structure

	Low	Hight
Low	(27.44,27.44)	(54.18,14.84)
High	(14.84,54.18)	(7.79,7.79)

Domestic Collusion Structure

	Low	High
Low	(28.20,28.20)	(55.67,15.17)
High	(15.17,55.67)	(7.97,7.97)

Collusion Structure

	Low	Hight
Low	(28.42,28.42)	(56.10,15.32)
High	(15.32,56.10)	(8.05,8.05)

Competition Structure

	Low	High
Low	(27.00,27.00)	(53.32,14.45)
High	(14.45,53.32)	(7.58,7.58)

The counterfactural experiments with different assumptions on the market structure. These market assumptions are used to back out the marginal costs similar to what implemented at Section 3.4.2. The top-left panel is for competition between two domestic firms and single product firms for the imported cars. The next top panel shows the results when marginal costs are derived by the assumption that the two domestic firms collude while importers competes. The down panels shows the results for the case that all firms compete or collude respectively. The numbers are in Million $ profit.

.3 Multinomial Logit Estimation

Table 10: Multinomial Logit Estimation

	Peykan	Vanet	RD	GLI	206 V2	Pars	Samand	LX	206 V3	206 V5	206 V6	405	Cielo
Sex	-3.58***	-4.37***	-2.94***	-3.01***	-1.46***	-2.57***	-2.66***	-2.42***	-1.44***	-1.60***	-1.79***	-2.68***	-2.19***
	-0.12	-0.36	-0.09	-0.27	-0.06	-0.11	-0.09	-0.15	-0.08	-0.10	-0.15	-0.07	-0.32
Family size	-0.04***	-0.12***	-0.11***	-0.13***	-0.28***	-0.19***	-0.15***	-0.09***	-0.34***	-0.35***	-0.31***	-0.14***	-0.12
	-0.01	-0.03	-0.01	-0.04	-0.02	-0.02	-0.01	-0.03	-0.02	-0.03	-0.04	-0.01	-0.06
Age	-0.01***	-0.03***	-0.01***	-0.01**	-0.02***	-0.02***	-0.02***	-0.01***	-0.02***	-0.02***	-0.02***	-0.01***	-0.03***
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
Education	0.01***	0.00	0.01***	0.01***	0.02***	0.02***	0.01***	0.02***	0.01***	0.02***	0.02***	0.02***	0.02***
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Income	-2.39**	-0.40	-1.10	-1.24	1.14*	1.15*	0.28	0.77	0.98	1.01	1.46	1.80***	0.47
	-0.81	-1.37	-0.73	-2.13	-0.46	-0.56	-0.60	-0.93	-0.71	-0.82	-0.88	-0.27	-2.31
Constant	-4.23***	-4.55***	-4.05***	-6.27***	-4.04***	-4.41***	-4.05***	-5.80***	-4.35***	-4.74***	-5.51***	-4.02***	-6.87***
	-0.09	-0.18	-0.09	-0.25	-0.10	-0.12	-0.09	-0.18	-0.14	-0.16	-0.22	-0.08	-0.41
	Mazda	Pajero	Renault	Maxima	Pickup	Patrol	Proton	Musso	Sinad	Verna	Gol	Merc.	Roniz
Sex	-1.99***	-2.31*	-0.95***	-3.26***	-46.53	-2.30*	-3.43***	-1.77***	-46.41	-1.03***	-1.12***	-2.64***	-2.01***
	-0.21	-1.05	-0.09	-0.42	-153.00	-1.05	-1.02	-0.41	0.00	-0.26	-0.26	-0.60	-0.34
Family size	-0.24***	-0.05	-0.35***	-0.05	-0.45***	-0.42	-0.35**	0.00	-0.11	-0.03	-0.17*	-0.12	-0.18*
	-0.05	-0.22	-0.03	-0.05	-0.09	-0.22	-0.13	-0.09	-0.25	-0.07	-0.08	-0.10	-0.08
Age	-0.02***	-0.08*	-0.01***	-0.02***	-0.02**	-0.03	-0.03*	-0.05***	0.00	-0.02*	-0.03**	-0.03**	-0.04***
	-0.01	-0.04	0.00	-0.01	-0.01	-0.02	-0.01	-0.01	-0.03	-0.01	-0.01	-0.01	-0.01
Education	0.02***	0.02	0.02***	0.02***	0.01	0.01	0.03***	0.01*	0.03	0.02***	0.02***	0.01	0.03***
	0.00	-0.02	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	0.00	-0.01	0.00
Income	2.12**	-11.44	1.33	1.29	1.36	-3.96	-0.36	1.59	-21.17	-3.13	-4.65	-0.04	1.93
	-0.77	-17.94	-0.75	-1.36	-2.30	-14.39	-5.48	-1.84	-23.82	-4.74	-4.91	-4.46	-1.16
Constant	-6.14***	-7.01***	-5.32***	-6.90***	-5.93***	-7.81***	-7.62***	-7.41***	-10.53***	-8.07***	-7.23***	-7.26***	-7.26***
	-0.30	-1.36	-0.18	-0.34	-0.51	-1.29	-0.73	-0.64	-1.69	-0.50	-0.47	-0.65	-0.45
	ELX	Nasim	Saba	Xantia	Rio	Caravan	Matiz	Prado	Corolla	Camry	Avante	Truck Mazda	
Sex	-2.58***	-1.85***	-2.14***	-2.77***	-1.69*	-2.17***	-0.95***	-3.30**	-2.84**	-2.74***	-2.04***	-3.70***	
	-0.39	-0.11	-0.03	-0.19	-0.77	-0.53	-0.20	-1.02	-1.03	-0.73	-0.53	-0.38	
Family size	-0.13*	-0.16***	-0.23***	-0.18***	-0.52*	-0.10	-0.22***	-0.29*	0.04	0.04	-0.38**	-0.17***	
	-0.07	-0.02	-0.01	-0.03	-0.21	-0.10	-0.06	-0.14	-0.14	-0.11	-0.12	-0.04	
Age	-0.02*	-0.01***	-0.01***	-0.02***	-0.02	-0.03*	-0.02*	-0.06**	-0.04	-0.06**	-0.01	-0.03***	
	-0.01	0.00	0.00	0.00	-0.02	-0.01	-0.01	-0.02	-0.02	-0.02	-0.01	0.00	
Education	0.02***	0.02***	0.01***	0.02***	0.02	0.01	0.02***	0.02	0.01	0.01	0.01	0.01**	
	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	0.00	
Income	0.40	-0.14	0.26	1.09	2.79	1.33	1.89	2.04	2.56	1.58	2.02	-1.58	
	-2.60	-1.13	-0.29	-0.86	-2.09	-2.73	-1.12	-1.82	-1.66	-2.14	-2.31	-2.36	
Constant	-7.32***	-5.34***	-2.33***	-5.24***	-8.25***	-7.63***	-7.20***	-6.48***	-8.48***	-7.40***	-7.73***	-5.32***	
	-0.43	-0.16	-0.04	-0.19	-1.15	-0.71	-0.39	-0.83	-1.04	-0.77	-0.73	-0.26	

Symbols ***, ** and * mean significance at 1%, 5% and 10%, respectively. Sex, age and education belong to owners' attributes, whereas Fimily size and income are family characteristics. Pars ELX, Peykan, Peykan Vanet, Peugeot RD, Peugeot GLI, Peugeot 206 V2, Peugeot Pars, Samand, Samand LX, Peugeot 206 V3, Peugeot 206 V5, Peugeot 206 V6, Peugeot 405 are produced by Iran Khodro Company. Renault, Maxima, Nasim, Saba, Xantia are manufacrued by Saipa-Yadak Company. Bahman Group Company produces Proton. Zagros Khodro Company imports Cielo, Matiz, Verna, Avante, Gol. The rest are imported by smaller companies. We designate the cars produced by I.K.Co, Saipa and Bahman as domestic cars and the rest as imported cars. *Merc.* stands for the car *Mercedes chairman*. In addition *Pride Nasim* and *Pride Saba* are full names of *Nasim* and *Saba*, respectively.

Bibliography

[1] Mark Aguiar and Gita Gopinath. Business cycles in emerging economies. the cycle is the trend. *Journal of Political Economy*, 115(1):69–102, 2007.

[2] David K Backus, Patrick J Kehoe, and Finn E Kydland. Dynamics of the trade balance and the terms of trade: The j-curve? *American Economic Review*, 84(1):84–103, March 1994.

[3] A.N. Berger, N.H. Miller, M.A. Petersen, R.G. Rajan, and J.C. Stein. Does function follow organizational form? evidence from the lending practices of large and small banks. *Journal of Financial Economics*, 76(2):237–269, 2005.

[4] S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

[5] S.T. Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262, 1994.

[6] Roberto Chang and Andrs Fernandez. On the sources of aggregate fluctuations in emerging economies. NBER Working Papers 15938, National Bureau of Economic Research, Inc, 2010.

[7] V. V. Chari, Patrick J. Kehoe, and Ellen R. McGrattan. Business cycle accounting. *Econometrica*, 75(3):781–836, 05 2007.

[8] Harold L. Cole and Lee E. Ohanian. New deal policies and the persistence of the great depression: A general equilibrium analysis. *Journal of Political Economy*, 112(4):779–816, August 2004.

[9] T.A. Domencich, D. McFadden, et al. *Urban travel demand: a behavioral analysis: a Charles River Associates research study.* North-Holland Pub. Co., 1975.

[10] Javier Garcia-Cicco, Roberto Pancrazi, and Martn Uribe. Real business cycles in emerging countries? *American Economic Review*, 100(5):2510–31, 2010.

[11] P.K. Goldberg. Product differentiation and oligopoly in international markets: The case of the us automobile industry. *Econometrica: Journal of the Econometric Society*, pages 891–951, 1995.

[12] Jeremy Greenwood, Zvi Hercowitz, and Gregory W Huffman. Investment, capacity utilization, and the real business cycle. *American Economic Review*, 78(3):402–17, June 1988.

[13] P.M. Guadagni and J.D.C. Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, pages 203–238, 1983.

[14] Constantino Hevia. Emerging market fluctuations: what makes the difference? Policy Research Working Paper Series 4897, The World Bank, 2009.

[15] K. Ho and J. Ishii. Location and competition in retail banking. *International Journal of Industrial Organization*, 2010.

[16] J. Hughes and L. Mester. Who said large banks don't experience scale economies? evidence from a risk-return-driven cost function. 2011.

[17] J.P. Hughes, W.W. Lang, L.J. Mester, and C.G. Moon. The dollars and sense of bank consolidation. *journal of banking & finance*, 23(2):291–324, 1999.

[18] J.P. Hughes, L.J. Mester, and C.G. Moon. Are scale economies in banking elusive or illusive?: Evidence obtained by incorporating capital structure and risk-taking into models of bank production. *Journal of Banking & Finance*, 25(12):2169–2208, 2001.

[19] R. Koenker. *Quantile regression*, volume 38. Cambridge Univ Pr, 2005.

[20] Finn E. Kydland and Edward C. Prescott. Time to build and aggregate fluctuations. *Econometrica*, pages 1345–1370, 1982.

[21] Ruy Lama. Accounting for output drops in latin america. *Review of Economic Dynamics*, 2010.

[22] L. Lambertini. Choosing rroles in an duopoly for endogenously differentiated products. *Australian Economic Papers*, 35(67):205–224, 1996.

[23] J. Levinsohn and A. Petrin. Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317, 2003.

[24] Ellen R. McGrattan. Solving the stochastic growth model with a finite element method. *Journal of Economic Dynamics and Control*, 20(1-3):19–42, 1996.

[25] L.J. Mester. Optimal industrial structure in banking. *Handbook of Financial Intermediation and Banking, ed. Arnoud Boot and Anjan Thakor*, 2008.

[26] Pablo A. Neumeyer and Fabrizio Perri. Business cycles in emerging economies: the role of interest rates. *Journal of Monetary Economics*, 52(2):345–380, March 2005.

[27] A. Nevo. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548, 2000.

[28] A. Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

[29] G.S. Olley and A. Pakes. The dynamics of productivity in the telecommunications equipment industry. *Econometrica: Journal of the Econometric Society*, pages 1263–1297, 1996.

[30] B.P. Pesek. Bank's supply function and the equilibrium quantity of money. *The Canadian Journal of Economics/Revue canadienne d'Economique*, 3(3):357–385, 1970.

[31] A. Petrin. Quantifying the benefits of new products: The case of the minivan. *Journal of political economy*, 110(4):705–729, 2002.

[32] M.H. Rahmati and S.R. Yousefi. Demand estimation for the iranian automobile industry. *Quarterly Review of Economics and Finance*, forthcoming.

[33] S. Rebelo and C.A. Vegh. Real effects of exchange-rate-based stabilization: an analysis of competing theories, 1995.

[34] Stephanie Schmitt-Grohe and Martin Uribe. Closing small open economy models. *Journal of International Economics*, 61(1):163–185, October 2003.

[35] Stephanie Schmitt-Grohe and Martin Uribe. Solving dynamic general equilibrium models using a second-order approximation to the policy function. *Journal of Economic Dynamics and Control*, 28(4):755–775, January 2004.

[36] C.W. Sealey and J.T. Lindley. Inputs, outputs, and a theory of production and cost at depository financial institutions. *The Journal of Finance*, 32(4):1251–1266, 1977.

[37] Ina Simonovska and Ludvig Soderling. Business cycle accounting for chile. IMF Working Papers 08/61, International Monetary Fund, March 2008.

[38] R.M. Solow. Technical change and the aggregate production function. *The review of Economics and Statistics*, 39(3):312–320, 1957.

[39] J.C. Stein. Information production and capital allocation: Decentralized versus hierarchical firms. *The Journal of Finance*, 57(5):1891–1921, 2002.

[40] Roman Sustek. Monetary business cycle accounting. *Review of Economic Dynamics*, 2010.

[41] J. Tirole. *The theory of industrial organization*. MIT press, 1988.

[42] Martin Uribe and Vivian Z. Yue. Country spreads and emerging countries: Who drives whom? *Journal of International Economics*, 69(1):6–36, June 2006.